UNIVERSITÉ CÔTE D'AZUR

Inria informatics mathematics

UNIVERSITY COTE D'AZUR

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Thèse de doctorat

Présentée en vue de l'obtention du grade de

docteur en Automatique et Traitement du Signal et des Images

de l'UNIVERSITE COTE D'AZUR

par Liuyun DUAN

Modélisation géométrique de scènes urbaines par imagerie satellitaire

Dirigée par Florent LAFARGE

Soutenue le 21st Avril, 2017

Devant le jury composé de :

Florence TUPIN	- Telecom ParisTech	*Rapporteur*
Przemyslaw MUSIALSKI	- TU Wien	*Rapporteur*
Florent LAFARGE	- INRIA Sophia Antipolis	*Directeur de thèse*
Mathieu DESBRUN	- California Institute of Technology	*Examinateur*
Frédo DURAND	- Massachusetts Institute of Technology	*Examinateur*
Lionel LAURORE	- Luxcarta Group	*Examinateur*
Justin HYLAND	- Luxcarta Group	*Invité*

UNIVERSITY COTE D'AZUR

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

P H D T H E S I S

to obtain the title of

PhD of Science

of the Université Côte d'Azur

Specialty : Automatique et Traitement du Signal et des Images

By Liuyun Duan

Geometric modeling of urban scenes from satellite imagery

Thesis Advisors : Florent Lafarge

prepared at INRIA Sophia Antipolis, Titane Team

defended on April 21st, 2017

Jury :

Reviewers :	Florence Tupin	-	Telecom ParisTech
	Przemyslaw Musialski	-	TU Wien
Advisors :	Florent Lafarge	-	INRIA Sophia Antipolis
Examinators :	Mathieu Desbrun	-	California Institute of Technology
	Fredo Durand	-	Massachusetts Institute of Technology
	Lionel Laurore	-	Luxcarta Group
Invited :	Justin Hyland	-	Luxcarta Group

Acknowledgments

I would like to express my sincere thanks to my reviewers : Professor Florence Tupin and Professor Przemyslaw Musialski for reading my thesis and giving insightful comments. I thank the committee members : Professor Mathieu Desbrun, Professor Frédo Durant, Mr. Lionel Laurore and Mr. Justin Hyland for attending as my examiners.

I would like to sincerely thank my advisor Florent Lafarge for his warm-hearted concerns of my adaptation to a completely new environment in France. I truly appreciated his patience and constructive advice for my research over these past three years of graduate studies. This PhD experience at the INRIA opened a great new world for me. I am grateful to all the help and training I received.

I would like to thank all my colleagues in Titane team : Pierre Alliez, Yuliya Tarabalka, Yannick Verdié, Kaimo Hu, David Bommes, Renata do Rego, David Salinas, Manish Mandad, Simon Giraudot, Sven Oesau, Jean-Dominique Favreau, Jean-Philippe Bauchet, Emmanuel Maggiori, and Hao Fang. I enjoyed very much discussing and solving problems with you. Many thanks to Florence Barbara, she helped me to solve all the administrative problems. Thanks for all the colleagues in ABS and Datashape teams, we shared a nice period of time exploring different interesting research works in our *after tea* events.

Tremendously many thanks have to be devoted to Luxcarta Group for funding my PhD and giving me the opportunity to further explore such an interesting topic. It is always a great honor to work for Luxcarta in such an impressive research atmosphere. I would like to thank Lionel Laurore, Justin Hyland, Véronique Poujade and Frédéric Trastour for all your technical discussions and support all the time. To all the colleagues in Luxcarta, I enjoyed a lot working with you and cherish all the time we shared.

I thank my family who loves me unconditionally. My decisions are always fully respected and understood. I am lucky to have the greatest parents and brother. I would like to thank all my friends. You cared for everything about me and brought me so much fun and pleasure.

Résumé

La modélisation automatique de villes à partir d'images satellites est l'un des principaux défis en lien avec la reconstruction urbaine. Son objectif est de représenter des villes en 3D de manière suffisamment compacte et précise. Elle trouve son application dans divers domaines, qui vont de la planification urbaine aux télécommunications, en passant par la gestion des catastrophes. L'imagerie satellite offre plusieurs avantages sur l'imagerie aérienne classique, tels qu'un faible coût d'acquisition, une couverture mondiale et une bonne fréquence de passage au-dessus des sites visités. Elle impose toutefois un certain nombre de contraintes techniques. Les méthodes existantes ne permettent que la synthèse de DSM (Digital Surface Models), dont la précision est parfois inégale.

Cette dissertation décrit une méthode entièrement automatique pour la production de modèles 3D compacts, précis et répondant à une sémantique particulière, à partir de deux images satellites en stéréo. Cette méthode repose sur deux concepts principaux. D'une part, la description géométrique des objets et leur assimilation à des catégories génériques sont effectuées simultanément, conférant ainsi une certaine robustesse face aux occlusions partielles ainsi qu'à la faible qualité des images. D'autre part, la méthode opère à une échelle géométrique très basse, ce qui permet la préservation de la forme des objets, avec finalement, une plus grande efficacité et un meilleur passage à l'échelle.

Pour générer des régions élémentaires, un algorithme de partitionnement de l'image en polygones convexes est présenté. Basé sur le diagramme de Voronoï, cet algorithme exhibe les régularités géométriques au sein des images, en alignant ces polygones le long de segments préalablement détectés. Inspirés par des travaux récents sur la reconstruction de scènes en 3D à l'aide de sémantiques particulières ainsi que sur la stéréoscopie, la géométrie et la classe sémantique des objets sont simultanément extraites des images, dans un processus commun de reconstruction et de classification. Les classes sémantiques radiométriques, les élévations et la forme géométrique des objets sont traitées simultanément au sein de chaque paire d'images. Il en résulte une plus grande robustesse face à la dégradation de la qualité de l'image

ainsi qu'aux aires d'occlusion partielle dans les vues satellites. Nos résultats prouvent le passage à la robustesse, l'échelle et la rapidité de l'approche proposée.

Mots clés : reconstruction 3D, modélisation de villes, imagerie satellite, scène urbaine, stéréoscopie, partitionnement d'images, classification sémantique, optimisation de contours, minimisation d'énergies, vision par ordinateur, géométrie computationnelle

Abstract

Automatic city modeling from satellite imagery is one of the biggest challenges in urban reconstruction. The ultimate goal is to produce compact and accurate 3D city models that benefit many application fields such as urban planning, telecommunications and disaster management. Compared to aerial acquisition, satellite imagery provides appealing advantages such as low acquisition cost, worldwide coverage and high collection frequency. However, satellite context also imposes a set of technical constraints as a lower pixel resolution and a wider that challenge 3D city reconstruction.

In this PhD thesis, we present a set of methodological tools for generating compact, semantically-aware and geometrically accurate 3D city models from stereo pairs of satellite images. The proposed pipeline relies on two key ingredients. First, geometry and semantics are retrieved simultaneously providing robust handling of occlusion areas and low image quality. Second, it operates at the scale of geometric atomic regions which allows the shape of urban objects to be well preserved, with a gain in scalability and efficiency.

Images are first decomposed into convex polygons that capture geometric details via a Voronoi diagram. Semantic classes, elevations, and 3D geometric shapes are then retrieved in a joint classification and reconstruction process operating on polygons. Experimental results on various cities around the world show the robustness, scalability and efficiency of our proposed approach.

Keywords: 3D reconstruction, city modeling, satellite imagery, urban scene, stereovision, image partitioning, semantic classification, contour optimization, energy minimization, computer vision, computational geometry

Acronym List:

ASA (Achievable Segmentation Accuracy)

BSDS500 (Berkeley Segmentation Data Set and Benchmarks 500)

CGAL (Computational Geometry Algorithms Library)

CRF (Conditional Random Field)

DEM (Digital Elevation Models)

DSM (Digital Surface Models)

DTM (Digital Terrian Models)

ERS (Entropy Rate Superpixel)

GCP (Ground Control Point)

GDAL (Geospatial Data Abstraction Library)

GIS (Geographic Information System)

IPM (Inverse procedural modeling)

LiDAR (Light Detection and Ranging)

LOD (Levels of Detail)

LSD (Line Segment Detector)

MRF (Markov Random Field)

MVS (Multi-view Stereo)

RPC (Rational Polynomial Coefficients)

SEEDS (Superpixels Extracted via Energy-Driven Sampling)

SGM (Semi-global Matching)

SLIC (Simple Linear Iterative Clustering)

VHR (Very High Resolution)

Contents

Chapter 1

Introduction

3D digitized urban models have a number of far-reaching applications. In the entertainment industry, for example, movies or video games whose storylines take place in real cities, require vivid and believable urban models created fully or partially from 3D modeling techniques. Likewise, the need for 3D city models in personal and vehicle navigation and location-based services is continuously growing. At a larger scale, urban planning, operational and training requirements for defense and emergency management, simulation of signal propagation for telecommunication planning, etc, benefit from virtual urban worlds.

Geometric modeling of urban scenes has received significant attention in recent years, especially in the broad field of computational science and computer vision. City reconstruction is an exciting research area with several valuable applications. A significant amount of previous work explored city modeling problems from various data sources, including terrestrial, aerial and satellite imagery, and Light Detection and Ranging (LiDAR). Still there remain many unsolved problems, particularly with regard to automatic city reconstruction from satellite imagery.

The goal of this thesis is to automatically reconstruct the 3D city model of a large-scale urban scene from satellite imagery. Due to the massive amount of available data and the limited resolution of satellite images, a proper solution should be scalable, robust to low image quality, and efficient in both computational memory and time. An automatic pipeline is proposed for producing compact and geometrically accurate large-scale city modeling from satellite imagery, which is able to robustly handle occlusions and low image quality by retrieving semantics and geometry simultaneously. Contrary to pixel-based methods, the proposed approach operates at the scale of geometric atomic regions: it allows the shape of urban objects

to be better preserved, and gains scalability and efficiency. Figure 1.1 illustrates the goal.

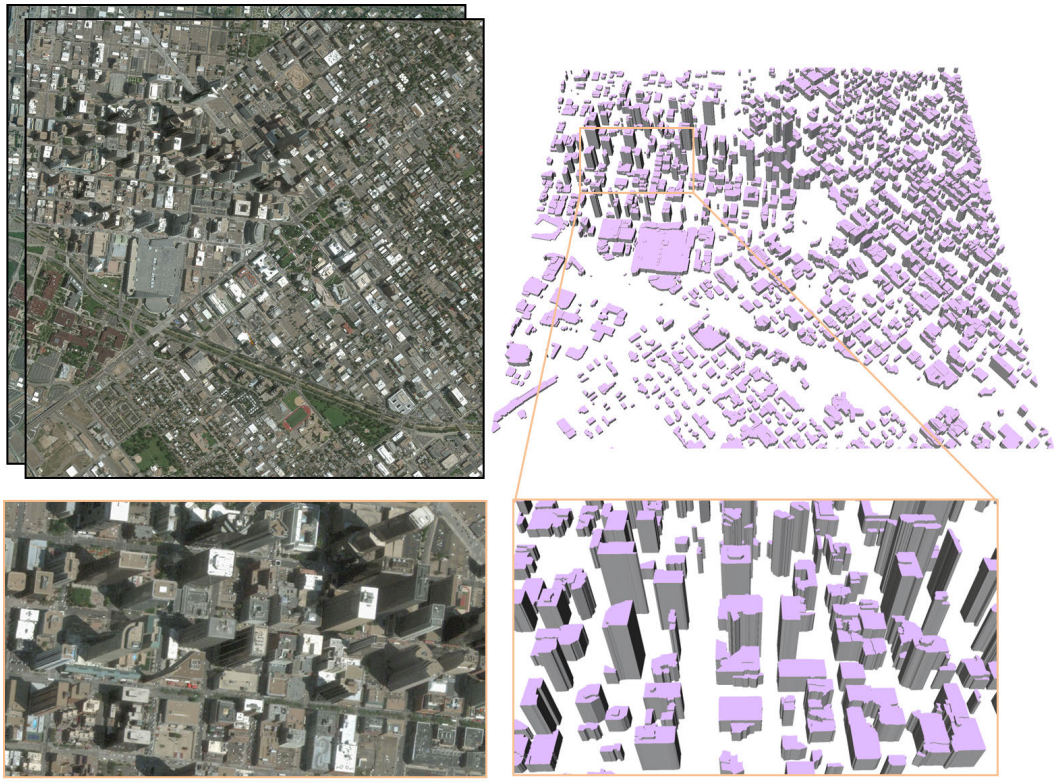


Figure 1.1: Reconstruction of Denver downtown. Starting from a stereo pair of satellite images (left), the automatic pipeline proposed in this thesis produces a compact and semantically-aware 3D model (right) in a few minutes.

1.1 3D city modeling: motivations and challenges

3D city models are represented as digital and usually georeferenced models of a city, which normally includes buildings, roads, vegetation and terrain as the most important feature types. Generally, the data sources consist of satellite imagery, aerial imagery, terrestrial imagery, airborne LiDAR and ground LiDAR.

Depending on the requirements of targeted applications, the output 3D city models are represented with different Levels of Detail (LOD) following the CityGML formalism [GP12], which distinguishes the geometric description and the specifica-

tion of the semantics and topology information as well:

LOD0: 2D footprints of objects are represented as vertical projections. Ground mesh is not used in this planar representation. Buildings are depicted as sets of polylines indicating the locations, and trees are marked as discs computed as virtual projection of tree models.

LOD1: A LOD0-building elevated in 3D with horizontal primitives as roofs (flat roofs, vertical facades). The heights are computed by median of heights in corresponding roof sections. Ground mesh is represented and trees are depicted as vertical cylinders. Details of roof structures or textures are omitted.

LOD2: Building models are reconstructed with piecewise-planar roofs (coarse roofs, vertical facades). Vegetation may also be represented in a geometric format, such as tree icons. No textures are mapped.

LOD3: Architecturally detailed building models including object semantics such as detailed roofs, facades, windows, etc. High-resolution textures can be mapped. Detailed vegetation and transportation objects can be represented in the model.

LOD4: Supplemented with indoor details like rooms, furnitures, doors etc., compared to LOD3 building models.

In recent years, the data acquisition technologies have evolved continuously. As the rapid increase of resolution of sensors, the line between LiDAR data and optical imagery is getting thinner. Dense (and semi-dense) point clouds can now also be generated from photogrammetry techniques. Depth cameras also provide possibilities to acquire extra information with additional hardware that measures geometric information about the sensors, such as accelerometers or GPS for geographic location, and compasses or gyrometers for orientation [Laf14].

These developments in data acquisition provides higher resolution and richer information for feature extraction. It allows practitioners to measure urban environments at the scale of individual shapes. However, the challenge lies in dealing with complex scenes which are composed of multiple objects, individually seen as an association of shapes. It requires the analysis of structural relationships between objects and shapes to understand principles of organization in urban scenes.

Based on the targeted applications, existing works are focused on different urban

scales, such as remotely sensed, terrestrial and indoor scenes. At the remotely sensed scale, descriptions of the main objects in urban environments are required. For instance, building roofs and road networks are expected for most of these applications. Driven by the progress of sensors in terms of quality and mobility, methodologies for feature extraction at terrestrial and indoor scales are well developed and now routine, providing data to reconstruct and analyze street views, facades, or inside scenes of rooms, etc.

The main challenges of city modeling lie in several aspects: acquisition constraints, quality of models, and full automation [MWA+13].

Acquisition constraints. The acquisition process is a generally difficult task in urban context due to the diverse defects which may come from any operation involved. Noise is one of the typical defects, and it can result directly from the sensor or from the data registration process. Trivial details, e.g., residual sensor noise, undesired vegetation, and vehicles, are the main sources of clutter in LiDAR point clouds. For laser sensors, the density of the points decreases according to the distance to objects. Depth information obtained through dense matching algorithms bring noise due to the ambiguities of pixel values, radiometric noise and weak textures. The most critical challenge for image data, is the missing parts caused by frequent occlusions or failures of matching, especially in complex scenes. Different data acquisition methods are discussed in Section 1.2.

Quality of Models. Quality of models involves different measurements and criteria. A *good* quality result is not only a model with a high geometric accuracy, or matching well to the physical scene, but also a compact representation with low complexity, and guarantees on the structural regularities on the outputs. Since the output model displays a city scene, a visual assessment is usually considered as a quality criteria as well. Multiple metrics need to be evaluated and some of them even conflict with each other. Defining a flexible and efficient strategy to combine different metrics becomes a challenge. Despite of the difficulty of evaluation criteria, the geometric accuracy is another big challenge for large-scale city scene reconstruc-

tion, particularly in a fully automatic way. From the point of view of interactive computer graphics, the accuracy of fully automatic reconstruction is quite low due to the ambiguity of data and the complexity of city scenes, while manual quality control suffers from the lack of scalability and efficiency if the amount of input data is huge.

Full automation. To be as automatic as possible is one of the ultimate goals for geometric modeling. Interactive modeling methods produce models with high accuracy and are often recommended for historical heritages or monuments while the number of objects is not massive. For large-scale scenes, the cost of human resources and time is not negligible for interactive modeling, while full automation is difficult to reach. Urban environments are generally complex with a great diversity of objects. Objects, even within a small scene area, are usually significantly different with respect to their density, complexity, and variety. Hence, pre-defined assumptions can hardly match the whole city, and vice versa, approaches with less assumptions are more flexible to variable applications but are less accurate, particularly in terms of object structures. On the other hand, algorithms with less assumptions usually involve complex and huge optimization tasks, considering the diversity in the scene. No unique solution is discovered for this paradoxical dilemma in automatic systems. Scientists turn to explore approaches which keep a balance in assumptions and accuracy, even with a tolerance of small interactions of human correction during modeling processes.

As with some contributions in the field, this thesis works on full automation for reconstructing LOD1 urban cities from satellite imagery. The proposed methodology takes a stereo-pair of satellite images as input, and produces the city scene model with all buildings represented as blocks, in a semantically-aware, scalable, flexible and time-efficient way.

1.2 Data acquisition

Sources suitable for 3D modeling include many types of data such as point clouds and optical imagery. Point clouds are usually generated from laser scanning or Multi-view Stereo (MVS) techniques. The precision of generated point clouds varies from different acquisitions such as LiDAR, Kinect and MVS with a webcam. Optical imagery is perhaps the most common and easy-acquirable data source. Imagery data from the ground captures local environments and individual objects with usually very high resolution and is flexible to obtain, to store and to exchange. With nearly ubiquitous Internet access, photos can be easily taken by smartphones and shared online. Many of these images describe urban scenes and are available to the public for urban reconstruction [PVGV+04, SSS06, IZB07, SGSS08, ASSS10, FFGG+10]. This data type is suitable to reconstruct local scenes, e.g., a few buildings [SSS+08], historical landmarks [IZB07] and street view facade reconstruction [AAC+06, HDT+07]. With the development of Web-mapping applications, such as Google Maps and Bing Maps, satellite and aerial imagery become available and widely utilized.

In this section, three representative types of data acquisitions in terms of decreasing resolution are discussed: LiDAR data, aerial imagery, and satellite imagery. Additionally, a brief introduction of Digital Surface Model (DSM) generation from images is given.

1.2.1 LiDAR data

Many applications have employed LiDAR data to represent the landscape topology, for instance of ocean floors, forest canopies, urban scenarios and mountain terrains. Laser pulses are directed on surfaces of objects and the reflected backscattering is captured, from which the structures and location information are computed through the principle of *time-of-flight* [CW11]. LiDAR data is usually very precise despite residual sensor noise, while contains some trivial details such as undesired objects. Different laser scanners are used with varying specifications according to the purpose, the size of the targeting area, the range of measurement and the cost

of devices. For different targets, in terms of materials, relevant wavelengths of wave signals are applied, varying from 250 nanometers to approximately 10 micrometers. In particular for detecting water surfaces, 532 nanometers is a typical choice, while 1064 nanometers is a default choice for urban acquisition, taking into consideration the absorbtion and backscattering characteristics of the target material. Generally, the smaller the target objects are, the shorter the wavelength that is applied. The limitation of LiDAR lies in its acquisition constraints: clear atmosphere conditions, daytime or nighttime coverage. Two of the main LiDAR data types are introduced: airborne LiDAR and terrestrial LiDAR.

Airborne LiDAR. 3D point cloud of a landscape can be generated from Airborne LiDAR using a scanner mounted on a plane or drone. The density of the point clouds is typically in a range of 1 to 50 points per square meter, with a low sampling error at a few centimeters. Until recently, airborne LiDAR has been considered as the most detailed and accurate data source for generating digital elevation models (DEMs) compared with photogrammetry, for its precision and high resolution. Due to the range of wavelengths, airborne LiDAR is able to filter out reflections from vegetation to create a DSM in which rivers, roads, buildings, etc., are surrounded by trees. This big advantage reduces the complexity of semantic classification of airborne LiDAR with respect to optical photogrammetry.

Terrestrial LiDAR. Terrestrial LiDAR scans scenes from a laser mounted on a supporting platform on the ground, e.g., a car, a tripod or a backpack. Data acquired in this way can be both stationary or mobile. Stationary terrestrial scanning is commonly utilized in applications of conventional topography, monitoring, cultural heritage documentation and forensics, as mentioned in [FZ04]. The density is usually higher than airborne LiDAR, generally in a range of 100 to 3000 point per square meter, and the sampling error is at the scale of millimeters. Texture information can be mapped from the digital images taken of the scanned area from the scanner's location. The color of each point in the point cloud comes from the corresponding pixel in the texture images. Thus, a visually realistic 3D model can be reconstructed

in a relatively short time, compared to other data sources.

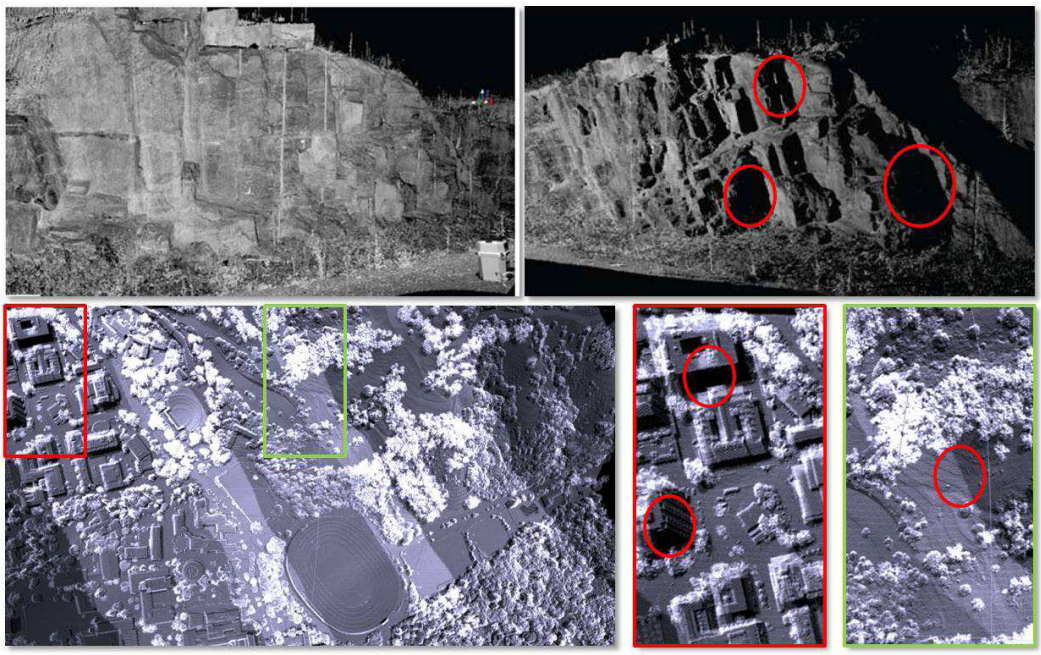


Figure 1.2: Examples of LiDAR data. The top row shows the terrestrial laser scanner data from two viewpoints, images in courtesy of [LDH10]. The bottom row shows the airborne LiDAR data collected in five flight passes with two closer crops on the right, image in courtesy of [OpenTopography13]. Note that red circles mark the holes and the strong discontinuities.

1.2.2 Aerial imagery

Aerial imagery has become a popular type of data acquisition for 3D city modeling, especially for urban city reconstruction with requirements of high accuracy and LOD levels. The aerial images are captured by cameras mounted on a plane or a drone. Depending on the size of area of interest, the needed resolution and the cost, different configurations can be employed. For example, a drone controlled remotely can be used to capture closer up views than a plane. The acquired aerial images for 3D modeling usually fit some standards, i.e., 80 percentage overlap in strip direction and 50 percentage overlap of strips. In order to recover the complete 3D information

from 2D projections in images, generally more than two images are required to cover the entire object. Aerial imagery has an obvious advantage in terms of the generally high spatial resolution coupled with radiometric texture information in the visible and/or non-visible wavelengths. However, an authorization is generally required to fly over the area of interest, and the acquisition cost is high. Aerial imagery can be collected in two typical modes: oblique aerial imagery and vertical aerial imagery that are usually integrated to acquire richer information both on roofs and facades.

Oblique aerial imagery. This concept refers to those images taken at a perspective angle from the air. According to the perspective angle, images taken from a low angle earth surface plane are called low oblique; otherwise those from a high angle are called high or steep oblique. MVS algorithms or other dense matching algorithms can be applied to generate 3D point clouds from a set of consecutive aerial images in different perspectives.

Vertical aerial imagery. Vertical aerial imagery is taken with a camera whose axis is directed towards the ground as vertically as possible. It is usually calibrated and mainly used as image interpretations for documentation. As supplementary information for 3D reconstruction applications, vertical aerial imagery can also be used to extract footprints of buildings, since it reduces the influence of shadows and occlusions from perspectives, and provides more accurate georeferencing.

1.2.3 Satellite imagery

Commercial satellite imagery has become widely accessible. It has received significant attention in the computer vision community. Satellites can be equipped with diverse sensors including acoustic, radar, ultrasonic or imaging sensors. Depending on the applications, specific sensors are applied to acquire relevant signals. Spatial resolution is often used as a major measurement to describe the details of information we can expect from a satellite image. It refers to the satellite sensor's ability to effectively capture a portion of the surface on the Earth (m^2) in a single pixel, typically varying from 30m to 0.3m. In this thesis, satellite images from

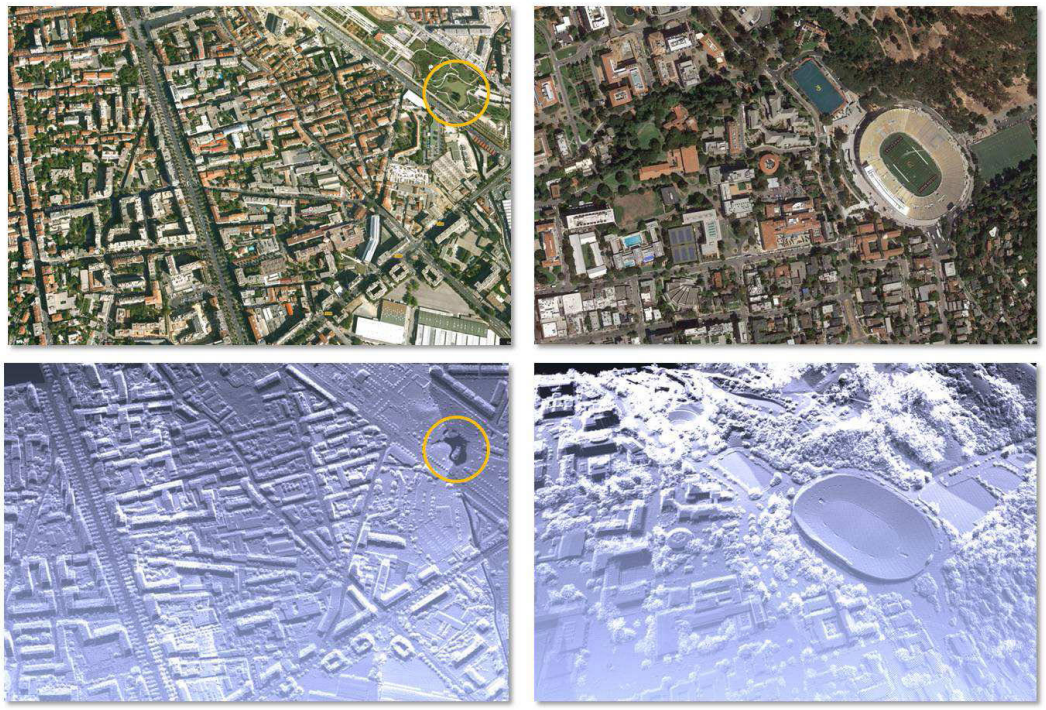


Figure 1.3: Aerial imagery. The first row shows the aerial pictures of Marseille, France, and the University of California campus in Berkeley, California, USA. The second row indicates the corresponding LiDAR data. Image courtesy of [Ver13]. Notice the gold circles in the left column indicate a "hole" in the LiDAR data.

QuickBird 2, WorldView 2, and Pleiades, with spatial resolution at 0.6, 0.46, and 0.5 meter respectively, are applied in the city reconstruction experiments. A brief introduction of the fundamentals of satellite stereo vision is given here, followed by a summary of advantages and disadvantages of satellite imagery for 3D modeling.

Fundamentals of satellite stereo vision. A camera model is fundamental for recovering 3D scenes from 2D images. Due to the complexity of rigorously modeling the physical camera model, as well as for Intellectual Property and security considerations, the Rational Polynomial Coefficient (RPC) model [HS97] is commonly used as an alternative sensor model to approximate the physical camera models. The RPC models describe the relationship between the underlying 3D scene geometry and its 2D image projections and is generalized for many applications. Furthermore,

for most satellite stereo analysis, geometric rectifications of images are required. The ultimate goal is to correct the camera models such that the rectified images match the epipolar geometry: that is, the corresponding pixels in the epipolar pair are located along the same row. This important characteristic benefits significantly dense matching algorithms by reducing the searching space from 2D to 1D [Hir08].

• ***RPC models.*** In practice, it is complicated to define a precise physical camera model due to the complex hardware configuration and the acquisition process. A representative approach to approximate the physical sensor models is the RPC model, which successfully achieves a high approximation accuracy while keeping the independence of sensors with a real-time calculation [HTC04]. RPC models fully use the physical parameters from the satellite, including focal length, principal point location, pixel size, lens distortions, and orientation parameters of the image such as position and attitude. Image pixel coordinates (Line, Sample) are formed as the ratios of polynomials of ground coordinates (Latitude, Longitude, Height). The coefficients can be determined by fitting the rigorous geometric model, with a set of given images and usually the aid of ground control points to minimize the fitting error. Once the RPC model is known, the mappings of image-to-object-space and object-to-image-space can be computed for a variety of computer vision and photogrammetry applications [Hir08, OTGB11, PC13, ZWDF15, GH15]. However, the physical sensor models can vary significantly across different satellites. Usually different models will be applied for the calibration of relative satellites in industrial production chains. [ZWDF15] proposes generalized sensor models to overcome this limitation of specificity and to achieve generalization across satellites. For exhaustive discussion about calibration and self-calibration, we refer the readers to [HS97, HZ04, MVGV09].

• ***Epipolar geometry.*** The camera model is simplified to the pinhole camera, for a standardization reason [MWA+13]. For one single camera, we don't have enough information to determine all the three parameters of an arbitrary 3D point projected to the image plane, but only two parameters to indicate the projecting

ray the point X lies on. The epipolar geometry is shown in Figure 1.4. All points on the projective ray from the first camera through x_1 in $image1$ and a 3D point X appears as a line l_2 in the second camera in $image2$. The further information we need to eventually locate point X is usually obtained by searching along the epipolar line l_2 for the matching projection x_2 of x_1. The line connecting both camera centers C_1, C_2 is referred as *baseline*, and the plane defined by C_1, C_2, X is referred as *epipolarplane*. The geometric rectification is a transformation of each image such that pairs of conjugate epipolar lines become collinear and parallel to the horizontal axis. Once the stereo pair is epipolar rectified, the corresponding pixel lies on the same row in the other image, which provides high efficiency for dense matching algorithms on large-scale satellite stereo images. In practice, epipolar lines are usually approximated by straight lines if the camera movement is almost linear, while epipolar lines are generally hyperbolas [GH97b]. When the image coverage is small, the influence of the error caused by the straight-line-approximation is negligible. The general steps mainly involve the detection of epipolar lines, and the rotation of the coordinate systems so that the epipolar lines are parallel to the horizontal axis.

Advantages and disadvantages. Compared with LiDAR and aerial imagery, satellite imagery has many advantages in terms of acquisition flexibility, coverage area, efficiency and cost. For example, the latest very high resolution satellite World-View 3, revisits the region within 1 day at a 0.31m ground sample distance. In most cases, the image data can be provided within 24 hours after the initial data has been acquired. LiDAR data and aerial imagery, on the other hand, are generally not available with such high revisit frequency and often require collection authorizations. Cost is an important consideration for many industries. The cost of satellite imagery over the same region is almost one tenth of LiDAR or aerial imagery. However, for urban city reconstruction from satellite imagery, there exist several challenges and limitations. The limited resolution of satellite imagery is a major hurdle for reconstructing 3D city models with high accuracy and level of detail. Occlusion problems brought by the constraint of long baseline, together with

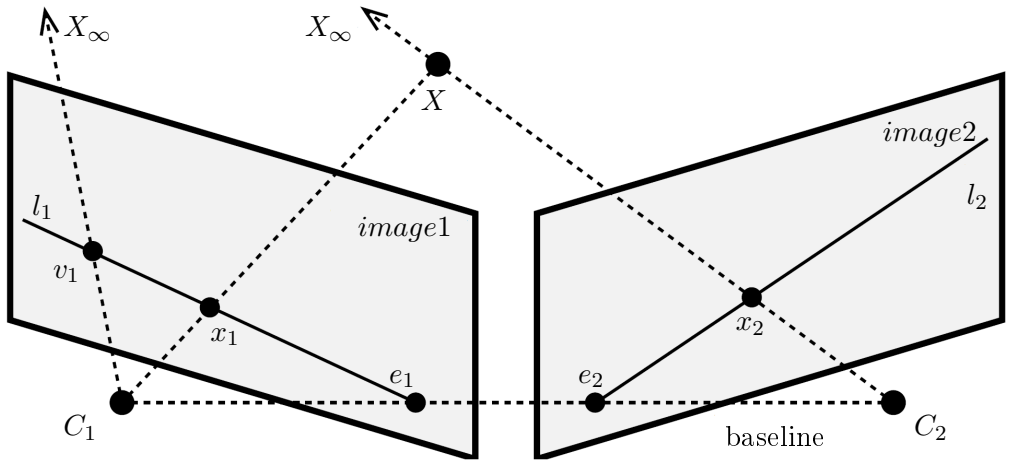


Figure 1.4: Epipolar geometry in a nutshell: points x_1 and x_2 are corresponding projections of the 3d point X. In $image1$ the point x_1 lies on the epipolar line l_1. The epipoles e_1 and e_2 indicate the positions where C_1 and C_2 project respectively. The point v_1 in $image1$ is the vanishing point of the projecting ray of x_2. Image in courtesy of [MWA+13].

the ambiguity of radiometry, make the computation of precise depth information difficult. In addition, atmospheric and cloud interactions sometimes influence the quality of satellite imagery as well.

1.2.4 DSM generation

For imagery data, dense matching is an important technique to generate DSMs. 3D shapes are estimated through two or more perspective views of the same scene.

Dense matching from two views, such as a stereo pair of satellite images, usually requires a geometric rectification of the images to fit the epipolar geometry. Then the remaining problem is to find the corresponding point in the other image. Global approaches like [KZ01, BVZ01] by graph cut, and [SZS03] by belief propagation, are quite memory intensive and not efficient enough for large-scale city modeling from satellite images. On the other hand, local methods are mostly based on correlation which can be implemented efficiently. However, these methods assume a constant

disparity inside a correlation window, and cause blurred object boundaries. To overcome these problems, Hirschmuller proposes a semi-global way to obtain both accuracy and efficiency named SGM [Hir08]. The SGM performs efficient and accurate stereo matching and is robust to recording or illumination variations. The process is based on a radiometric robust matching cost and an optimization with a global smoothness constraint. It approximates a global cost by piece-wise matching based on the epipolar geometry and a global smoothness constraint. It can be performed in a linear time to the number of pixels and disparities and is almost as accurate as global methods. Due to the high accuracy and efficiency, the SGM is valuable for creating models of large-scale urban scenes.

For dense matching from more than two views, for instance aerial photogrammetry, [Col96] first proposed a *plane-sweeping* process to determine the 2D feature correspondences and the 3D positions of feature points in the scene. It works for true multi-image matching with an arbitrary number of images n, in a linear algorithmic complexity with n. For stereo matching from pushbroom images, De Franchis demonstrates a stereo pipeline to produce digital elevation maps automatically from Pléiades satellite images provides by CNES (the French Space Agency) [DF15]. On the other hand, to handle the large massive aerial images, many methods are proposed to improve the efficiency, such as [FP10, AHL15, XGL16] based on patch-based multi-view stereo (PMVS), [Hir08, RWFH12, HCW+16] based on the SGM. Furthermore, [YCZ+16] proposes an algorithm based on optical flow field, which achieves good matching efficiency and flexibility for complex terrain surface matching.

Popular dense matching algorithms such as PMVS [FP10], MicMac [PDP06], semi-global-block-matching (SGBM) [Its15], SURE [RWFH12] and S2P[C. De Franchis], are publicly available for academic researches.

1.3 Related work

The related work of 3D city modeling mainly involve problems in categories of: scene classification, urban object extraction, and 3D reconstruction approaches.

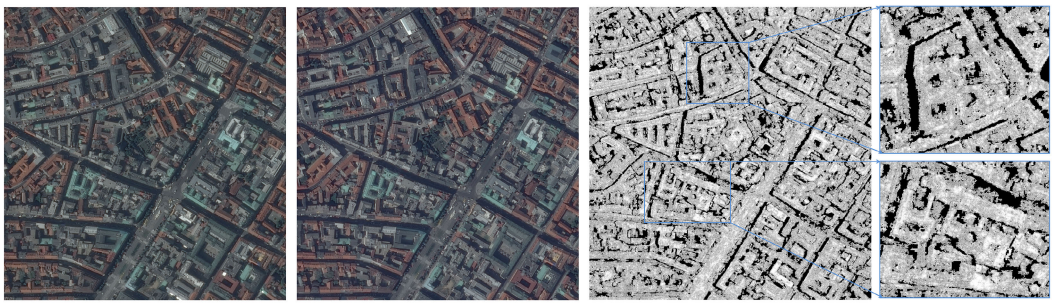


Figure 1.5: Dense matching from satellite imagery. From the left to the right are respectively the two-view images of the stereo pair in epipolar geometry over Prague, and the disparity map computed by the SGM (thanks to Dr. Heiko Hirschmüller).

1.3.1 Scene classification

Many different classification algorithms have been explored in photogrammetry and computer vision domain. The goal is to partition the entire scene into different classes of interest. Scene classification approaches can be categorized into several groups : supervised/unsupervised, local/nonlocal based, and contextual/non-contextual aware.

Supervised/unsupervised methods. Classification approaches in this group are the most widespread. Supervised classification requires prior definition of class signatures through a learning procedure, with a set of representative samples for each land cover class. Unsupervised classification generally groups pixels by clustering local feature descriptions into target classes, without learning involved. Methods such as [SJC08], directly apply Texton Boost and Texton Forest to the input data without using any feature descriptor, which performs a high computation efficiency by avoiding filter-bank response. Furthermore, [SWRC09] learns a discriminative model of object classes integrating texture, layout, and context information, and applies texton-based features in texture-layout filters. [LSR+12] focuses on classifying challenging street view data, optimizing the problems of object classification and dense stereo reconstruction simultaneously. This method dramatically improves the classification accuracy to 95.8% in the street view data set for Leuven. [GKF09]

applies learning algorithms to large point clouds, exhibiting high accuracy even in the presence of residual sensor noise and undesired vegetation and vehicles. The limitation is the requirement of a training process to learn the shapes of each different building type or object. [HZC+13] combines the learned image-based appearance likelihoods and geometric priors, and depth maps. The interaction of these clues yields an improved dense reconstruction and labeling. Unsupervised approaches do not require learned classifiers, but rely on feature descriptions such as spectral, textural, and geometrical attributes, intensity, and bag-of-visual-words (BOW). Many unsupervised approaches are based on graphical models [Bis06]. Smoothness constraint is used in [RvDHV06] to decompose the input data by searching smoothly connected regions. Classification and height maps are combined in [TTM07] to enhance the handling of buildings based on synthetic aperture radar (SAR) images. Different information is merged into a Markovian framework and the classification results are improved. A joint object segmentation and stereo matching is presented in [BRK+11], processing in an unsupervised mode. It is less accurate than supervised methods such as [LSR+12], but more general and flexible. Benefiting from the joint mechanism, this method works well for street view with complex environments including cars, trees, passengers and occlusions.

Local/nonlocal-based methods. Most local-based classification approaches take locally independent features such as intensity, texture, color to distinguish different data classes. For example, [SJC08, XQ09, ZWY10] perform the classification relying on pixel-based local features, without neighboring spatial constraints and concentrating only on the local appearance of specific classes. Other approaches, nonlocal-based, integrate local feature descriptors and spatial correspondence to enhance the classification performance. [MPM+14] proposes a segmentation method based on shape descriptors and spatial configurations of neighboring patches. [CJSW01] takes into account both feature space and spatial relation between pixels simultaneously. Similarly, [HTP05, KOSPK16] perform automatic segmentation of VHR satellite imagery by integrating texture, intensity, and color features within a pixel-wise refinement framework. Some of the most general meth-

ods for nonlocal based classification are graph-based optimization, e.g., Markov Random Fields (MRFs) and Conditional Random Fields (CRFs). [PTN09] proposes a hierarchical CRF, in which the higher level nodes describe the class-label configuration of the smaller regions, and the results from an image classifier are considered as global features. [MZWLS15] classifies urban areas in aerial images by a CRF with higher-order potentials to balance the data term with the object candidates.

Contextual/non-contextual methods. Contextual information provides clues of relevance to improve the semantic classification. [GRB08] models the co-occurrence and the relative locations using a CRF, considering both semantic and the spatial relevance using pairwise features. [LRKT13] models co-occurrence statistics though a global potential function that defined over all variables in the CRF, to measure the likelihood of the occurrence of classes in the same images together. [BRK+11] employs contextual constraint from object level, using both 2D radiometric consistency and 3D spatial connectivity to classify objects and reconstruct the 3D model. [FVS12] applies superpixels for simultaneous class segmentation and object localization over a CRF, using context and certainty of superpixel quality as constraints. [HZC+13] uses class-specific smoothness assumptions to improve the quality of the obtained reconstruction. [JY12] relies upon a message-passing algorithm to solve a holistic CRF that includes contextual terms such as a scene and likelihoods of class presence. [MCL+14] improves both semantic segmentation and object detection by involving global and local contexts. Some other classification approaches are not contextually-aware. For example, [MCA+16] uses appearance class probability and geometric class prior in a joint framework and classifies multiple classes by solving a sequence of convex optimizations while progressively removing constraints. This hierarchical scheme shows excellent improvements on processing time and memory efficiency.

1.3.2 Urban object extraction

Automatic extraction of urban objects such as vegetation, buildings or roads has important practical value for semantic detection and scene understanding. The

most interesting semantic objects in urban scene reconstruction are buildings, road networks and vegetation.

Footprint of buildings. Buildings are the most common observed objects in urban city scenes. A high quality building extraction is of significant importance for the generation of realistic 3D models of urban environments from imagery. Generic models for building extraction assume all buildings follow a certain pattern [RHM02]. [The06] proposes a prototype using the existing improved snake energy function to compute the snake contours, but initializes the model with the proposed circular casting algorithm instead of radial casting algorithm [MZC05]. [ZBKB08] extracts buildings from aerial imagery with dense height data and sparse 3D line segments, combined with a rough building mask estimated from the height data. This method is fully automatic and takes advantage of multiple data sources. [ST10] developed an approach for the automatic extraction of the rectangular and circular shaped buildings from high resolution satellite imagery using the Hough transform. In this method, candidate building patches are detected from the imagery using a SVM classifier with extra bands of information applied to enhance the classification. Shaker et al. [SAEAGS11] create DSMs from the IKONOS stereo imagery to extract building heights. Due the additional information, the experimental results show an average building detection percentage at 82.6% in dense residential areas. [AM12] proposed an object based approach by using stable and variable features together. Stable features are derived from inherent characteristics of building phenomenon and variable features are extracted using separability and thresholds analysis tool. [TQCR16] proposes a robust time-series data analysis method, using spatial-temporal information of the stereo imagery and the DSMs generated from them. It computes building probability maps to identify building objects.

Road networks. The detection of road networks provides not only traffic usabilities for city modeling, but also valuable instructive clue for building detection or reconstruction since these two semantics are highly contextual. Classification-based methods usually use the geometric features, photometric features and tex-

ture features of a road [HK92, TK03, Sim11, LC14, PZ14]. Because of the difficulty in distinguishing road and other spectrally similar objects such as building blocks, water areas and parking lots, etc., methods in this group have limitations in terms of accuracy. Knowledge-based methods apply parametric models, such as an energy function model, to extract road networks. The common parameter models usually extract structural elements according to the relationship among them, and to detect the specific structure to fit the goal of road network detection [SLS08, VCB10, MWL12, CFL13]. These methods have a better description of the structural shapes but suffer from the disadvantage of over-extraction and non-robustness to occlusions and shadows. Other approaches, such as [Zha04] combine the given GIS data and color information in the input stereo images, improving the robustness to occlusions and shadows. For a detailed survey of road network extraction, we refer the readers to [WYZ+16] for a more complete introduction.

Vegetation. The extraction of vegetation is important for the generation of realistic visualizations and useful for city planing. Many methods have been published to extract landscapes or trees from LiDAR point clouds [JA07, CA09, LM11]. These approaches usually employ both geometry and reflection properties and extract vegetation though general classifiers such as thresholding, SVM, Adaboost, etc. Mapping vegetation through remote sensing images involves considerations of image processing techniques. Some approaches extract vegetation based on spectral radiances in the red and near-infrared regions, incorporating the spectral vegetation indices (VI) [AFKH84, GDB85, BPJ07]. To further improve the ability to distinguish more detailed vegetation groups, [LL05] applied a more advanced image classification method by sub-pixel analysis. [CRZC04] chose to increase the resolution of the images in order to obtain better performance but with a higher processing cost. [TCGV09] proposes an approach based on spectral unmixing and statistically developed decision trees to classify urban vegetation characteristics. This method needs no assumptions with respect to the frequency distributions of the input image data.

1.3.3 3D reconstruction

The scientific literature related to 3D city modeling is massive across several communities in computer graphics, computer vision, photogrammetry and remote sensing, and robotics. Driven by different interests of virtual applications, considerable developments have been made during recent years such as approaches highlighted in [VAB10, TVG11, BSRG15, VLA15, PSP16]. Approaches for 3D city reconstruction use various strategies including: procedural modeling, inverse procedural modeling, mesh simplification, volumetric modeling, primitive-based modeling, superpixel-based modeling, and joint methodologies.

Procedural modeling. Procedural modeling algorithms work efficiently to create artificial 3D models in a fast and scalable way. The structure of object shapes is parametrized into a set of grammars. The model is generated by iteratively applying the rules on an initial shape (e.g., a box). Based on predefined grammars, these methods are powerful and efficient when applied to modeling 3D buildings [VAM+10, MWH+06]. It produces a lightweight semantically meaningful representation instead of massive meshes. Procedural modeling approaches usually consist of two types of strategies, shape grammar [Sti75] and split grammar [MWH+06] for constructing buildings and facades [MZWVG07, VAM+09, MWA+13]. The output models are compact, editable, readable, semantically-aware, advantageous for retrieval and fast graphics generation [HWA+10], and accurate due to the predetermined rules and procedural mechanisms [VAM+09]. However, procedural modeling requires writing detailed grammars, and there are many modeling scenarios that lack structural and parametric representations. Real city scenes contain diverse architectural styles and complex structures, and it is a difficult task to get all the rules and grammars predefined a priori. For the most part, procedural modeling is mainly used for creating geotypical models and not reconstructing real environments.

Inverse procedural modeling (IPM). Compared to procedural modeling methods, IPM approaches take shape grammars as a higher order knowledge for the reconstruction of scenes. The rules or/and grammars are extracted from given models

(images) of buildings, plants, cities and worlds to create a generative system. Hence it involves dealing with gigantic search spaces for the variety of objects in city scenes. The problem of efficient search is a critical challenge in the field. [Hig05] presents a comprehensive survey about grammatical inference, and the smallest context-free grammar problem is examined in [CLLS05]. These works show that it is an NP-complete problem compared with grammar-based compression algorithms [NMW97b, NMW97a]. In this part, two major groups of IPM algorithms are discussed: facade modeling methods and building modeling methods.

• ***Facade modeling.*** Driven by the applications of ground-based city navigation and realistic virtual cities in the entertainment industry, many works have been devoted to street level modeling from ground or oblique-view data. Accurate 3D facade models are produced by applying different strategies, e.g., example-based system [XFZ+09], general principle of Minimum Description Length (MDL) [TYK+12, MVG13, WYD+14], layered modeling [ZXJ+13, IMAW15], grid-based graph data structure [SHFH11, DCNM14], image parsing with machine learning [MMWVG12], and co-occurrence clustering [LW15]. These approaches analyze structural rules and grammars from either images or point clouds and automatically compute a useful procedural description as output. Lower level shape understanding, most importantly symmetry detection, is an important step for inverse procedural modeling from noisy input or input that is not segmented [WYD+14]. Feature lines are applied in [BBW+09] to detect symmetric parts without restrictions of regular patterns or nested hierarchies. In addition, interactive modeling combines manual editing and procedural modeling together [LWM08] to obtain precise shape grammars and realistic reconstruction. In general, facade modeling concentrates on relatively regular shapes (windows, doors etc.) that involves smaller search spaces for optimization and works well to create visually pleasing models.

• ***Building modeling.*** A large amount of work has been published for building modeling using IPM [PMW+08, VAB10, DAB14, LWWS15, PSP16]. Categorized by the type of input data, these building models can be produced from images,

meshes and point clouds. **(I.)** [ARB07] follows an image-based approach to recover a style grammar of geometric models from a sparse image set by subdividing buildings into feature regions and construct novel architectural structures in the original style. With the Manhattan-world assumption, [VAB10] defines a grammar for representing changes in building geometry. It reconstructs geometric 3D models from one or more calibrated aerial images associated with each building by an optimization of geometric and photometric matching. A limitation of this work is its non-robustness to windows and shadows. **(II.)** Mesh-based methods extract rules or/and grammars from 3D meshes which contain more precise and complete structural information compared to images. [BWS10] detects context-free grammars and presents a both theoretical and practical framework for IPM of 3D buildings. [DAB14] works on the proceduralization of buildings at city scale, converting existing 3D unstructured urban models into an editable and compact procedural representation. Grammars have been used as high level priors in [VGDA+12], which are extracted by a search with MCMC. **(III.)** Point-based methods take the advantage of rich information and high accuracy, but face the challenges of data completeness and noise [MGP06, TMT10, MMWV11, DAB15, LWWS15]. A matching scheme is presented in [LWWS15] to find consistent co-occurrence patterns in a frame-invariant way, which is robust in handling geometric variability (noise, irregular patterns, appearance variation, etc.). [DAB15] creates a tree representation of the detected repeating structures, and then generates the grammar of 3D buildings using a consensus-based voting scheme and a pattern extraction algorithm. Some approaches such as [TMT10] generate a sparse tree with planar patches as terminals, models are produced by merging terminals. [PSP16] uses the symmetry prior for convex variational 3D reconstruction, well suited for locally symmetric architectures even with noisy or incomplete data. **(IV.)** Further approaches can be more general in terms of input data type. [KBW+12] presents a theoretical framework for characterizing shapes by building blocks using r-similarity w.r.t. rigid transformation, suitable for synthetic data. A learning method is applied by [TMT16] to produce augmented procedural models with neural networks that learn how to satisfy constraints.

The IPM methods have in common that a critical problem is to convert existing scenes into procedural representations (rules or/and grammars). However this is a *chicken & egg* problem since we have no complete structural knowledge of the objects we are modeling [ADBW16]. In general, IPM methods work well on facade modeling but are not equally successful for the modeling of entire buildings. Because of the variety and complexity of urban building structures, it is hard to get the parameters of all possible types of complex architectures in an efficient way.

Mesh simplification. Differing from Procedural or Inverse Procedural Modeling, mesh simplification methods produce 3D models in a format of compact and smooth meshes instead of CAD models. The goal is to reduce the number of vertices and faces while keeping close to the input mesh. Brief reviews of mesh generation methods and mesh simplification approaches are given as follows.

• ***Generation of dense meshes*** The input mesh for mesh simplifications is usually generated from a set of points obtained from MVS techniques or LiDAR scanning. MVS generates dense point clouds from each pixel in images (w.r.t. sparse point clouds from matching feature points in SfM). As a typical technique in MVS, SfM extracts salient features from each registered images, and applies a clustering within a feature space. 2D tracks are constructed by matching these feature points between images, from which the position of cameras are automatically recovered by solving the SfM model [HZ04]. SfM provides networks of registered images, their camera properties, and sparse point cloud. Since MVS is sensitive to reprojection errors, a bundle adjustment is often required. For a detailed overview, we refer the reader to [SS02, SCD+06] for two-view stereo methods and multiview stereo methods respectively. With the developments of techniques such as camera calibration and dense matching, MVS produces high quality point clouds. Compared with LiDAR, it still often suffers from severe noise and null data holes [LKB10]. In terms of data density and accuracy, LiDAR is a powerful data source for mesh simplification methods.

•***Simplification approaches.*** From the input data, polygonal models are first generated directly by rasterization or Delaunay Triangulation, then simplified with general mesh simplification algorithms, usually involving edge collapse and vertex relocation by minimizing several error metrics [GH97a, LT99]. The accuracy measurement is obtained with two error metrics: the Accurate Measure of Quadratic Error (AMQE) and the Symmetric Measure of Quadratic Error (SMQE). The mesh simplification step significantly reduces the number of triangles while preserving a low fitting error [ZN10, SLA15]. An error-driven approach is presented by [CSAD04], which applies variational geometric partitioning to approximate the mesh with a set of shape proxies. [ZN10] proposes a 2.5D dual contouring method, using an adaptive grid data structure to reconstruct the geometry in each grid by minimizing a Quadratic Error Measure. This method produces 2.5D city models with geometry and topology created respectively by discovering global regularities. Mesh decimation is a powerful solution for mesh simplification. It provides an approach to greedily cluster geometric elements, creating a partition of the original input mesh. [SLA15] explores the decimation of triangle surface meshes by detecting a set of planar shape proxies and structuring them via an adjacency graph. The fidelity is guaranteed by the planar proxies, such that the simplification approach is robust to defect-laden meshes. In [LKB10], the proposed hierarchical method starts from an 3D-surface mesh obtained by MVS, and then inserts a set of detected primitives into the surface. It combines the mesh-based surface (describes details) and 3D primitives (describes regular shapes), hence it achieves a high compression ratio while keeping details. Due to the statistics analysis in primitive extraction, this strategy allows the introduction of semantic knowledge, the simplification of the modeling, and correction of registration errors.

Volumetric Modeling Volumetric modeling is a data-driven approach and has been proved to be a robust way of generating crack-free models [LC87, CL96, JLSW02, FBGS05]. Volumetric approaches partition the scene into elements such as cubic voxels [VTC07, GDDA13], tetrahedras [JP11, VLPK12] or polyhedral cells [CLP10]. These elements are classified as inside or outside. Volumetric methods

are known to produce non-smooth surfaces with watertight characteristics [LPK07]. However, due to the voxel-based mechanism, these methods generally suffer from poor scalability [JP11], but see a few applications to large-scale urban reconstruction from aerial images [CLP10, GDDA13]. [CL96] employs a Marching Cubes method [LC87], and integrates a large number of range images to produce seamless, high-detail triangle mesh models. Recently, [HKR+12, HKDB13] propose a way to apply the inside/outside reasoning as a new input from SfM and locally update the tetrahedralization. An updated mesh is finally obtained on-the-fly. [FBGS05] develops a modeling method combining aerial and ground-based LiDAR, based on the dual contouring method [JLSW02] which creates one mesh vertex in each minimal grid node by optimizing a quadratic error function and polygons are constructed by a traversal over the adaptive grid. On average, volumetric representations contain both inside and outside information, which is an advantage but also a limitation since it is difficult to model, time consuming and costly in terms of storage space, compared to mesh surfaces.

Primitive based. Primitive-based modeling methods are mostly based on the detection of predefined roof shapes, e.g., planar shapes [VKH06, MSS+08, PY09, ZN09, LWC+11], given user-defined primitives [YHNF03, LDZPD08, ZBKB08, ZN08]. These methods work well for the reconstruction of buildings composed of pre-defined shapes, while lose accuracy when dealing with arbitrary roof structures. Approaches of automatic pipelines [VKH06, MSS+08, ZN08, PY09, ZN09, ZN12], introduce a segmentation module separating points into ground and individual building patches. Vegetation and noise are removed in advance through a classification process. Then mesh models are constructed by modeling algorithms over these patches, and final buildings are reconstructed through different heuristics based on the extraction of planar shapes. The topology between planar roof patches can be extracted by a graph-based method [VKH06], and roof contours can be regularized by estimating building orientations [MSS+08]. With the segmented regions from a classification process, lightweight and watertight polygonal models are created by simplifying boundaries of fitted planes [PY09]. Roof boundaries are aligned by learning a set of

principal directions [ZN08], which is further extended to city scale data in [ZN09]. [LDZPD08] searches for the optimal combination of parametric models using a RJMCMC sampler [Gre95]. Furthermore, [LWC+11] applies RANSAC to detect primitives and uses a global mutual relations between these primitives to generate synthetic models. For interactive applications, [YHNF03] shows the reconstruction of complex building roofs and irregular shapes by refining the primitives with user interaction. In general, primitive-based methods produce reasonable structures that follow the user-defined primitive libraries, but are limited when handling arbitrary shapes.

Superpixel based. Some 3D modeling approaches are based on atomic regions (superpixels). For example in [PNF+08], it presents a real-time plane-sweep stereo on the GPU, using the orientations and plane priors detected from SfM point clouds. Further, [MK10] aggregates the superpixels of surface data and extends the plane sweeping to handle textureless regions. [FCSS09] presents an efficient modeling way by simplifying the detected planes over a Markov Random Field, to generate a piecewise planar Manhattan-world geometry. A semantic analysis is performed in [BSRG14] based on superpixels. This strategy is further explored by combining with image edges to perform per-view dense depth optimization [BSRG15], and produces meshes with competitive surface quality efficiently. In summary, superpixel-based modeling methods usually offer high scalability, time-efficiency, and capability to integrate semantics and geometry.

Joint reconstruction. Reconstruction of urban objects and scenes has been deeply explored in vision, with a request towards full automation, quality and scalability, and robustness to acquisition constraints [MWA+13]. Recently, many joint methods are proposed to combine different informations and take an advantage of their implicit relations. Thus these different clues mutually benefit from each other through an global optimization scheme, improving the reconstruction quality. Most of joint reconstruction methods either simultaneously explore semantics and geometry, or integrate multiple data acquisitions as input sources.

•***Semantically-aware approaches.*** In the urban reconstruction domain, geometry and semantics are closely related. The most traditional strategy consists in retrieving semantics before geometry. In many city modeling methods [LM11, VLA15], the input data is first segmented into different classes so that the subsequent 3D reconstruction can be adapted to the nature of urban objects. Recent works [HZC+13, LGZ+13, CSF15] demonstrate that the simultaneous extraction of geometry and semantics, also known as semantic 3D reconstruction, outclasses multiple step strategies in terms of output quality. Based on a Conditional Random Field (CRF) [LMP01] formulation, [LSR+12] introduces a learning based method to jointly optimize the semantic classification and the disparity assignment. Similarly, [BRK+11] develops a joint optimization for object class segmentation and depth estimation over a CRF, but in an unsupervised way with constraints of 2D consistency and 3D geometry preservation. However, these works typically suffer from a low scalability and often produce 3D models without structural consideration. Semantic 3D reconstruction remains a challenge at the scale of satellite images. In addition, to further integrate more semantics, [BSRG15] reconstructs the surface based on image edges, superpixels and second-order smoothness constraints. The geometry and topology are preserved by cooperating image discontinuities and 2D shape analysis. The mesh is simplified in a strategy of point clustering and region segmentation, which is comparable to classical MVS reconstruction.

•***Multiple data sources.*** Different data acquisitions present specific attributes and advantages. City reconstruction methods integrate multiple data sources, such as aerial data, ground data, and depth maps, to gain more perspective information and more clues from both 2D and 3D spaces. A method presented in [FSA05] creates building models with facade by integrating two acquisitions of airborne LiDAR and terrestrial LiDAR. [HSU06] similarly combines aerial images and ground images to achieve more detailed reconstruction. [ZBKB08] integrates 3D line segments and global optimization of dense matching technique to obtain a globally optimal level of details of depth information and presents a fully automatic reconstruction from aerial imagery. Based on the footprints of buildings extracted from optical images,

SAR information is integrated for the building presence validation and the height retrieval in [STD09]. Methods in this group employ more information and clues to reconstruct 3D urban scenes with higher accuracy and level of details, while increasing the complexity and producing cost of the system.

Conclusion. In summary, many approaches have been explored for 3D modeling of city scenes. Few of them, however, are scalable and adaptive to large-scale urban scene reconstruction from satellite imagery. How to handle the complex contexts found in satellite imagery in order to efficiently and accurately reconstruct semantic objects is still a challenge in the 3D modeling filed.

1.4 Satellite context

Satellite imagery imposes a set of technical constraints with respect to traditional aerial imagery, in particular (i) a lower spatial resolution, typically ≥ 0.5 meter, (ii) a lower signal-to-noise-ratio (SNR) impacting the image quality, and (iii) a wider baseline to guarantee a reasonable depth accuracy. Although these constraints have a low impact on some applications such as change detection [GH15] or generation of dense DSMs [ZWDF15], they challenge the automatic extraction of semantics and the reconstruction of semantically-aware city models.

Characteristics. Satellite images over urban scenes capture the city optical appearance from space. Usually, 3D information from satellite imagery is derived from overlapping stereo pairs. In order to preserve a reasonable height precision, the baseline of the stereo pair is necessarily long so that the H/B ratio (height/baseline) is relatively small [GFMP08]. However, a long baseline brings huge differences in terms of perspectives, which introduces obvious occlusions from objects. Especially in modern downtown areas, lower buildings close to skyscrapers may be severely occluded. Facades of the same buildings may be visible in one image, but completely missing in the other one. For buildings with non-flat roof structures (e.g., dome, gambrel, rhombic, conical, spire etc.), images from two perspectives are insufficient to recover complete 3D information since they are partially-self-occluded.

Urban contents such as buildings, roads and vegetation are disseminated in the entire scene with/without correspondence from each other. Particularly, buildings are constructed with diverse architectural structures, surrounded by various undesired objects. The appearance of roofs may differ significantly due to temporary objects on top. Vegetation can be problematic for building reconstruction since it introduces ambiguities on discontinuities and confusion with buildings.

Small buildings may be occluded or even embedded inside dense vegetation. Due to the limited spatial resolution of satellite imagery, it is difficult to precisely recover small buildings or small details of non-flat roofs for lack of information, not to mention noise or interfering surroundings. A typical satellite context is shown in Figure 1.6.

Challenges. Satellite imagery is interesting for 3D city modeling for its high acquisition flexibility, large area coverage, and relatively low cost, while imposes several technical challenges with respect to traditional aerial acquisitions.

• ***Image quality constraints.*** Optical satellite images are taken from a camera mounted on a satellite. Atmospheric conditions and cloud cover influence the visibility of the Earth's surface. Specific weather conditions such as fog and clouds appear commonly, which usually cause weak contrast in images. Additionally, artifacts caused by perturbations of the signal from sensor saturation or A/D (analog/digital) conversion, bring speckle-stripes or artificial structures. Depending on the type of sensor used, the sensor noise can be modeled as either additive, multiplicative (speckle) or impulsive (salt and-pepper) [GW08]. Furthermore, to recover the 3D information from 2D projections more than one image is required. Stereo images are rarely captured at the same moment, which brings ambiguities of the radiometric information of the scene, particularly from temporary objects.

• ***Depth calculation challenges.*** Remote-sensing data need an essential preprocessing for geometric corrections due to varying acquisition and transmission conditions. Based on the RPC models, images are calibrated to satisfy the epipo-

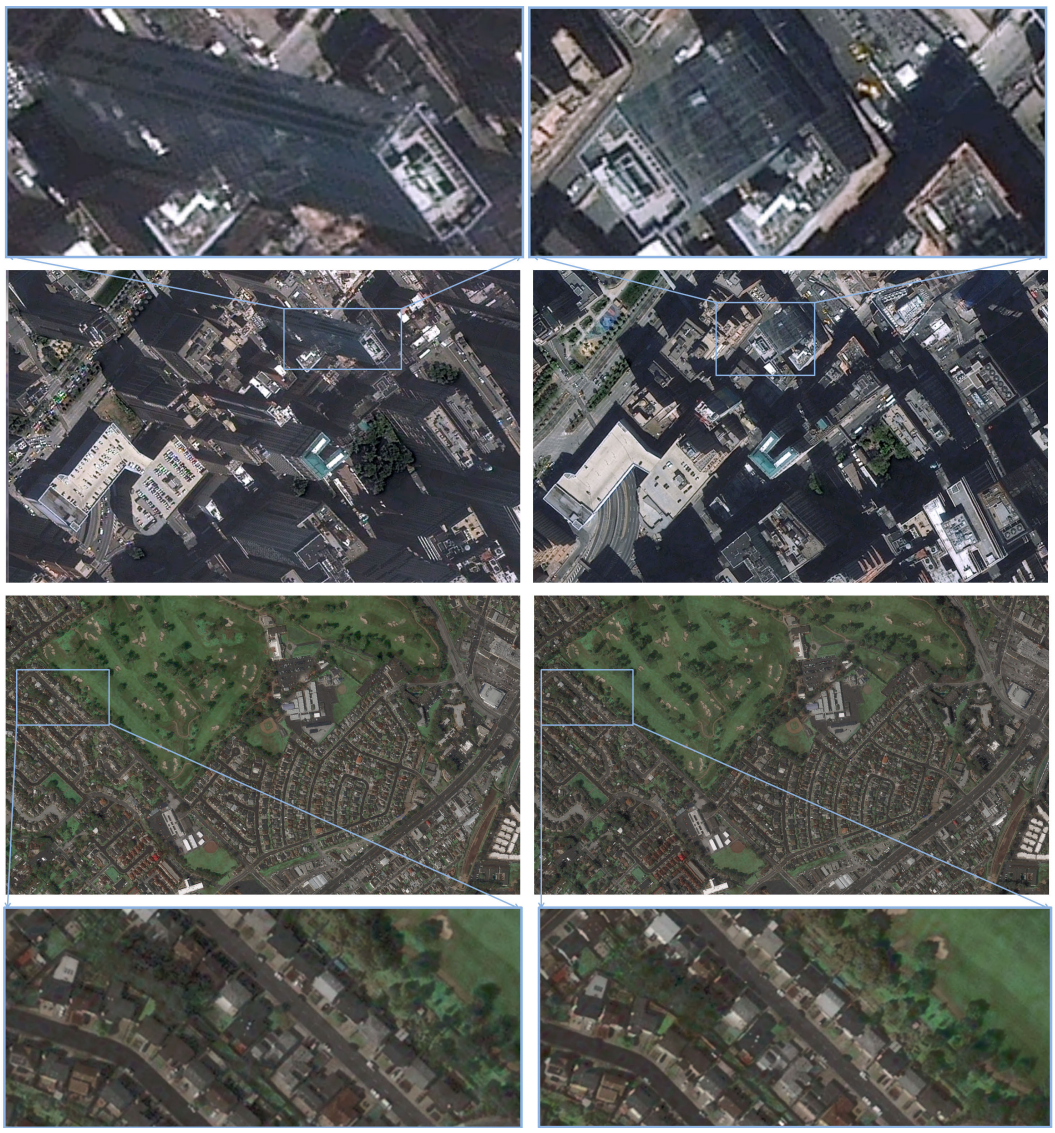


Figure 1.6: Satellite context. Top row: a wide baseline is a necessity to reach reasonable depth accuracy, but brings severe occlusion problems. A facade side is typically visible only in one image (see close-ups). Note also the high proportion of shadow and the time-varying objects as cars. Bottom row: dense residential areas bring challenges to precisely reconstruct each individual building from the non-negligible local variations in elevation (inset shows where houses in San Francisco, US, are built on high-relief terrain).

lar geometry with aids from a set of Ground Control Points (GCPs). However the precision of calibration significantly impacts the dense matching calculation. The tolerance of geometric error is usually required to be less than 1 pixel. On the other hand, the long baseline makes it sometimes impossible to compute the depth of each pixel due to occlusions. For example, the SGM searches for the matching pixel along the relative epipolar line in the other image, which works well for object boundaries and discontinuities but not for occluding parts, repetitive patterns and textureless regions. Another difficulty is the matching precision over slopes or domes whose depth map changes continuously and rapidly. As a consequence, disparity maps acquired from large baseline satellite images are usually sparse and noisy.

• ***Automatic 3D city reconstruction challenges.*** Aerial acquisitions with LiDAR scanning or multi-view imagery constitute the best way so far to automatically create 3D models on large-scale urban scenes [MWA+13]. Because of high acquisition costs and authorization constraints, aerial acquisitions are, however, restricted to some spotlighted cities in the world. In particular, Geographic Information System (GIS) companies propose catalogs with typically a few hundred cities in the world. Satellite imagery exhibits higher potential with lower costs, a worldwide coverage and a high acquisition frequency. Satellites, however, have several technical restrictions that prevent GIS practitioners from producing compact city models in an automatic way [PC13].

Large-scale city reconstruction from satellite imagery is particularly challenging due to lacking image quality and resolution. The spatial resolution of high quality satellite imagery is limited by the given sensor, which varies from approximately 0.31m to 1.5m. For instance, in a satellite image with 0.5m resolution, a $10m \times 10m$ roof reflects 20×20 pixels, which is usually not sufficient for applications that require a high accuracy in detail. Most methodological approaches usually separate the automatic city reconstruction from satellite imagery into several hierarchical tasks: image classification [CW11, MTCA16a, MTCA16c], semantic extraction [CPK+14, MTCA16b, WYZ+16], and specific object reconstruction through general 3D modeling methods as detailed introduced in Section 1.3.3. Existing methods pro-

duce at best dense Digital Surface Models at a varying degrees of fidelity. Automatic city modeling from satellite imagery is still one of the biggest challenges in urban reconstruction.

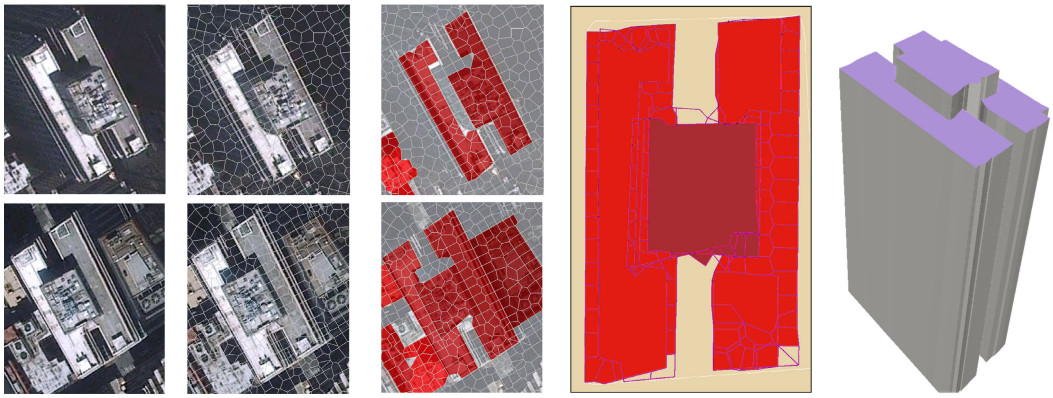
1.5 Contributions

Inspired by recent works on semantic 3D reconstruction and region-based stereovision, an automatic pipeline of urban city reconstruction from satellite stereo pairs is proposed for producing compact, semantically-aware and geometrically accurate 3D city models in an efficient way. The intuitions include two key ideas: first, geometry and semantics are retrieved simultaneously to handle occlusions and low image quality problems; second, the reconstruction operates at the scale of geometric atomic regions to gain scalability and efficiency, and preserve object shapes.

From a calibrated stereo pair of satellite imagery, the produced output city model is a compact mesh composed of ground and building objects. Buildings are represented with a LOD1 of the CityGML formalism [GP12], i.e., piecewise planar buildings with flat roofs and vertical facades. The automatic pipeline proceeds with three main steps illustrated in Figure 1.7.

In this thesis, Chapter 1 introduces the context of 3D city modeling and summarizes the related work on different objectives in this domain as well as the particular challenges with satellite imagery. In Chapter 2, the first step of the proposed pipeline, a novel polygonal partitioning algorithm is presented. It decomposes images into convex polygonal atomic regions, preserving the shapes of urban objects. In Chapter 3, the second step, a joint classification algorithm is proposed for classifying the atomic regions into semantic classes by simultaneously optimizing the problems of semantic labeling and elevation estimation. In Chapter 4, the third step, a model fusion approach is proposed for producing geometrically accurate object contours from the two preliminary models produced by the second step. Experiments of the reconstruction over different urban cities in various styles are presented in Chapter 5. Conclusions and perspectives are given in Chapter 6.

The main contributions are summarized as follows:



(a) (b) (c) (d) (e)

Figure 1.7: Overview. Input stereo images (a) are first decomposed into atomic convex polygons (b) using a polygonal partitioning algorithm (Chapter 2). In a second step detailed in Chapter 3, the semantic class and the elevation of each polygon are simultaneously retrieved in the two partitions (c). The last step (Chapter 4) consists in unifying the two partitions enriched by semantic classes and elevation values into a planimetric elevation representation (d) that allows the generation of the output 3D model (e).

(i) A fully automatic pipeline is proposed for producing compact and semantically-aware city models from satellite images, which is time-efficient, scalable, and able to reconstruct large cities in a few minutes.
(ii) A novel algorithm is proposed to partition images into convex polygons, operating at the scale of the geometric shape, and not directly at the pixel scale, which introduces several interesting properties in terms of geometric guarantees, region compactness and scalability.
(iii) A joint classification and reconstruction process that brings robustness to the low quality of input images, is proposed.

Chapter 2

Polygonal partitioning

In this chapter, a polygonal partitioning method is proposed for decomposing the image into convex connected polygons through a Voronoi diagram. The over-segmentation of satellite images into atomic regions is a powerful strategy to reduce computational complexity. Traditional superpixel methods, that operate at a pixel level, cannot directly capture the geometric information disseminated into the images. An alternative to these methods is proposed by operating at the level of geometric shapes. The overall strategy consists in building a Voronoi diagram that conforms to preliminarily detected line-segments, before homogenizing the partition by spatial point process distributed over the image gradient. This method is particularly adapted to images with strong geometric signatures, typically man-made objects and environments. Comparisons are conducted with state-of-the-art superpixel methods, and the potential of the proposed approach is demonstrated with experiments on large-scale images in Appendix 7.1.

2.1 Introduction

Partitioning images into meaningful atomic regions is very popular to address vision problems. When used as pre-processing for image segmentation [LZMC12], stereo matching [ZK07b] or object boundary extraction [LSD10] for instance, such an image decomposition offers very interesting advantages in terms of algorithmic complexity and spatial consistency. Traditional methods create image partitions at the pixel level, atomic regions being commonly called *superpixels*. Each region is delimited by a set of pixels forming a free-form contour. This representation brings high flexibility, but is free of higher level information. In particular, it does not

exploit geometric information disseminated into images. This is particularly penalizing in some applications or specific contexts for which the shape and adjacency of regions are expected to have strong geometric constraints. In stereo matching for instance, guaranteeing convexity of regions make the matching procedure more robust than the subsequent extraction of their convex hull [BSRG14]. Also, in presence of man-made objects and urban environments [RBF12, CDSHD13], preferring regions with straight line boundaries is a natural choice which can be a precious source of geometric knowledge for subsequent processing steps.

Few existing works have addressed the problem of the geometric partitioning of images. Integrating such geometric knowledge after superpixel decomposition is a complex and delicate task. Inconsistencies within the graph of region adjacency are frequent and lead to structural incoherences in subsequent processing. Also, modifying the region shapes typically destroys the effort done to make superpixels adherent to the image. Ideally, both geometry and radiometry must be jointly exploited to generate the regions.

In this chapter, a novel partitioning method is proposed for decomposing images into atomic regions with convex polygons, while imposing geometric guarantees on the shape and connection of these regions. Figure 2.1 illustrates our goal.

The proposed solution consists of partitioning images into connected convex polygons using Voronoi diagrams for which a brief introduction is given in Section 2.3. Region convexity has many advantages, in particular for (i) simplifying subsequent geometric operations as the computation of region distances, (ii) favoring the region compactness, and (iii) insuring a unique adjacency graph between regions, without ambiguities. In our approach, geometric properties are guaranteed by construction of the Voronoi diagram whereas radiometry is exploited to (i) align edges separating two neighboring polygons with image discontinuities, and (ii) center the polygons in homogeneous areas. Contrary to interest point based-strategies [Tuy10, CFL13], image discontinuities are approximated through the detection of geometric shapes, i.e., line-segments similarly to [SCF14, ZFW+12].

The proposed algorithm takes an image as input and produces, as output, a partition into polygons defined into the continuous bounded domain supporting the

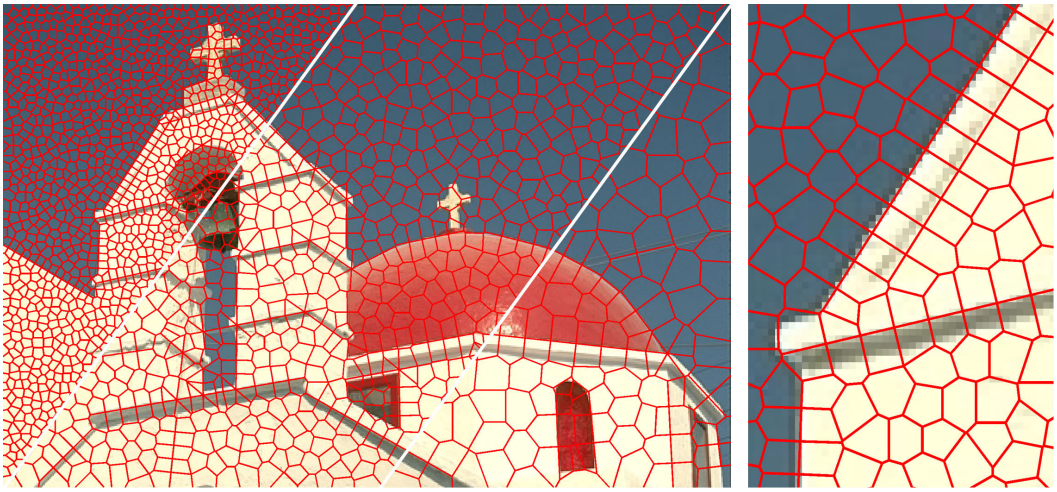


Figure 2.1: Goal of polygonal partitioning. Our algorithm partitions images into regular convex polygons. Three different polygon sizes are displayed. The use of floating polygons allow for the preservation of object boundaries at a subpixelic scale (close-up).

image. A model parameter ε has to be specified to fix the partition scale; concretely ε corresponds to the average radius of a region, assuming the region approaches a rounded shape. The main contributions are summarized as follows:

- *Shape anchoring.* A strategy to preserve geometric shapes within the Voronoi partitions is proposed. The key idea relies on the insertion of pairs of Voronoi seeds, called *anchors*, close enough to each other to constrain the Voronoi edges to be part of a geometric shape. Beyond preservation, the connexion of the geometric shapes are also structured, in particular for enhancing shape junctions.
- *Geometric guarantees.* The output provides some geometric guarantees related to the shape and adjacency of the atomic regions. First, each region is a convex polygon with a low number of edges. Contrary to many superpixel methods, the adjacency of regions is also guaranteed to be unique by construction, two polygons being neighbors if they share a common edge. Finally, region boundaries are polygons with exact geometry, i.e., under the pixel scale.
- *Efficiency.* By manipulating geometric entities, the pixel-based information

is simplified and the algorithmic complexity of the partitioning process is strongly reduced. If the efficiency of superpixel methods can be strongly affected by big size images, the proposed algorithm is weakly impacted both in terms of time efficiency and memory consumption.

The proposed strategy is composed of three steps illustrated in Figure 2.2.

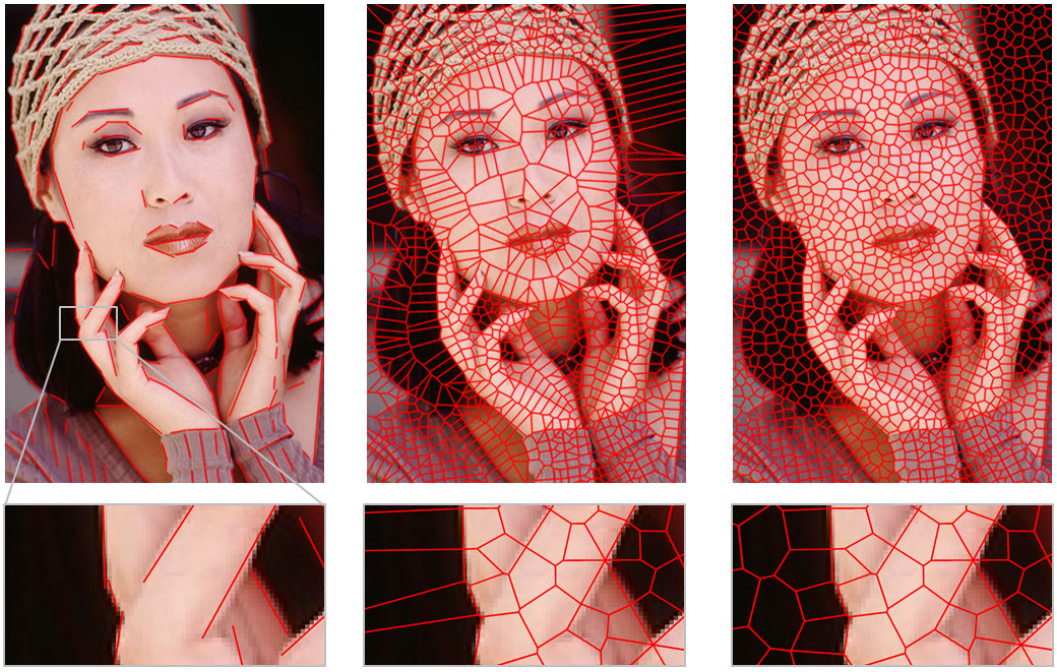


Figure 2.2: Overview of the polygonal partitioning. Left: line-segments are first extracted from the input image, and consolidated to bring spatial coherence (Section 2.4). Middle: an initial Voronoi partition that preserves the line-segments and their junctions is then created by inserting anchors at specific locations (Section 2.5). Right: the Voronoi partition is homogenized by point process (Section 2.6).

2.2 Review of image partitioning

The review of previous work related to image partitioning covers two main facets of our problem statement: segmentation into superpixels and shape detection.

Segmentation into superpixels. Methods partitioning images into superpix-

els are usually evaluated on five criteria: (i) adherence to boundaries, (ii) running time, (iii) compactness of regions, (iv) memory efficiency and (v) simplicity of use. Among the numerous algorithms proposed in the literature, the most popular strategy consists in iteratively refining superpixels from an initial rough partition of pixels. These methods, e.g., [ASS+12, LSK+09, VdBBR+12, WW12, ZWW+11], are usually time and memory efficient and capture boundaries well. Some methods address the problem with more global strategies, in particular with energy minimization on graphs [LTRC11, RJRO13]. Results are usually of higher quality but require more algorithmic efforts. Globally speaking, each method has its own advantages and drawbacks, and scores differently on the five criteria mentioned above. Nevertheless, adherence to boundaries is usually favored at the expense of region compactness by a large majority of methods [SFS12]. Apart from certain algorithms as SLIC [ASS+12], no control on the shape of regions is possible.

Geometric shape detection. The automated detection of geometric shapes is an instance of the general problem of fitting parametric functions to data. There is a wide variety of shapes in all dimensions, the most common one in image processing being line-segments. This parametric shape is known to capture well the image discontinuities, in particular for man-made environments. Interpreting line-segments from images can bring precious information for discovering the scene structure [LHK09] or recognizing people [Ren07]. If the Hough detector has been widely used in the literature, recent algorithms deeply improved the quality of line-segment detection while guaranteeing fast running times [DHH11], and even false detection control [VGJMR10]. Closely related to line-segments, textons [ZGWX05] also provide a compact representation of the image structure in between the pixel and geometric shape scales.

2.3 Mathematical background

Two mathematical tools that play a central role in our algorithm are briefly introduced: Spatial point processes and Voronoi Diagrams. Deeper presentations of these tools can be found in [BVL93, OBSC00].

Spatial point process. A point process describes random configurations of points $P = \{p_1, ..., p_n\}$ in a continuous bounded set K, in our case the 2D domain of the input image. The number of points n is itself a random variable that typically follows a discrete Poisson distribution. What makes point processes appealing is the possibility to create spatial interactions between points, in particular using the Markovian properties, i.e., points interact only in a local neighborhood. The most common process using Markovian interactions is the Strauss process in which a repulsion domain is located around each point to avoid points to be too close to each other. When $\dim K = 2$, this domain is a disk whose radius is a model parameter. The sampling of point processes is usually a fastidious operation relying on Monte Carlo methods [VL14]. However, fast sampling mechanisms exist for certain types of point processes. This is the case of Strauss processes for which efficient Poisson-disk sampling allows the random generation of points either homogeneously distributed [DH06], or following an arbitrary density [BSD09].

Voronoi diagram. Given a configuration of points P in K, called *seeds*, the Voronoi cell associated to the seed $p_i \in P$, denoted as $V(p_i)$, corresponds to the region in which the points are closer to p_i than to any other seed in P:

$$V(p_i) = \{x \in K \,/\, \|x - p_i\| \leq \|x - p_j\|,\ \forall p_j \in P, i \neq j\} \tag{2.1}$$

The Voronoi diagram generated by P is the set of the Voronoi cells $\{V(p_1), ...V(p_n)\}$. Voronoi digrams have appealing geometric properties, in particular they entirely partition the domain K without region overlap. By using the Euclidean distance in Eq. 2.1, Voronoi cells are guaranteed to be convex polygons. The dual graph of a Voronoi diagram also corresponds to the Delaunay triangulation of its seeds, and

gives the adjacency relation between regions. Finally the algorithmic complexity to build a Voronoi diagram when $\dim K = 2$ is only in $O(n \log n)$.

Spatial point processes can be used to generate the seeds of a Voronoi diagram. In particular, Poisson-disk sampling constitutes a fast and efficient way to create partitions of homogeneous Voronoi cells.

2.4 Shape detection

The first step of our algorithm consists of extracting line-segments from the input image, and then consolidating them to bring spatial coherence.

Line-segment extraction. As mentioned in Section 2.2, many methods have been proposed in the literature. Our choice focuses on the Line-Segment Detector (LSD) [VGJMR10] for the detection quality, the running times and the false detection control. The minimal length of line-segments is fixed to ε. Note that our algorithm is not restricted to LSD and can be used with other line-segment or polyline extraction methods.

Consolidation. The extraction of line-segments is a local process that can generate heap of shapes with noise and outliers. Such raw detected line-segments is sometimes hardly exploitable. Thus, a consolidation procedure is proposed to bring spatial coherence among the line-segments. An adjacency graph is built: two line-segments l_i and l_j are considered as adjacent if $d(l_i, l_j) \leq \varepsilon$, where $d(.,.)$ is the minimal euclidean distance between any pair of points of the two line-segments. As illustrated in Figure 2.3, sets of adjacent line-segments are consolidated using three types of operators:

- *Merging.* The merging operator tests whether two adjacent line-segments are near-collinear, and, if valid, replaces them by one large line-segment that covers their length.
- *Removing.* A small line-segment is removed when adjacent to a large near-parallel line-segment.

– *Concurrence.* The concurrence operator tests whether the inscribed circle of three mutually adjacent line-segments, i.e., of a simple cycle of order 3 in the adjacency graph, is small, and, if valid, translates the three line-segments onto the center of the inscribed circle.

Note the adjacency graph is updated after each effective operations. Merging, removing and concurrence operators are successively applied over the line-segments using a greedy procedure.

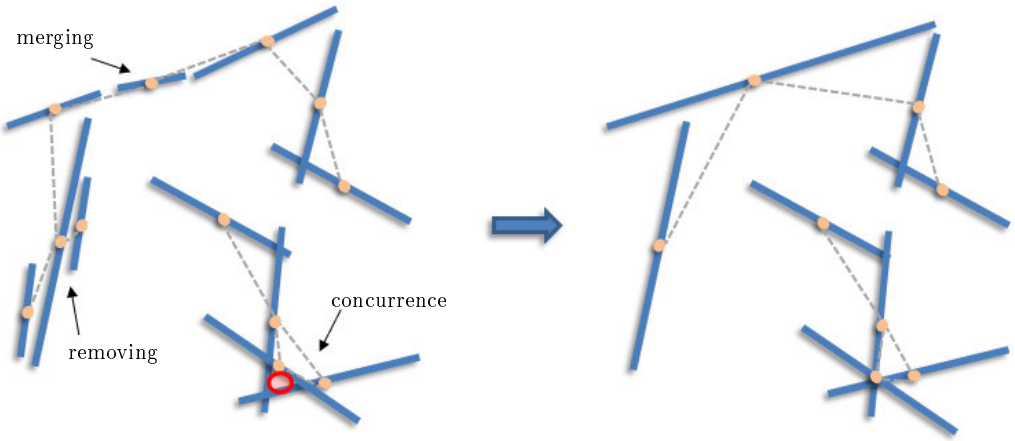


Figure 2.3: Line-segment consolidation. Three different operations applied greedily over the adjacency graph (dashed grey lines) bring spatial coherence between the detected line-segments. Such a consolidation procedure also reduces the problem complexity as the number of line-segments becomes lower.

2.5 Shape conforming Voronoi partition

Our objective is now to create a Voronoi partition that conforms to the detected line-segments. Said differently, the line-segments must not cross the Voronoi cells, but must be included onto the Voronoi edges. Manipulating a Voronoi partition to make the cells align with some geometric shapes or radiometric information is a delicate operation because, for the displacement of a single seed, even small, the whole group of connected cells is usually strongly perturbed. This explains why the use of Voronoi diagrams in vision has mainly be restricted to the creation of basic

isotropic partitions, e.g., in texture segmentation [TJL88]. Inspired by a recent work in surface reconstruction to conform 3D Delaunay triangulation to planes [LA13], a mechanism is proposed to create a Voronoi partition that conforms to the line-segments by construction.

Shape anchoring. The key idea consists in sampling pairs of seeds, called as *anchors*, located on each side of a line-segment. As illustrated on Figure 2.4, each anchor is positioned so that the Voronoi edge separating the cells induced by the two seeds is exactly on the line-segment. The two vertices of an anchor are reflectively symmetric to the line-segment distant by ε. The anchors related to the same line-segment are regularly positioned with a distance of 2ε. This strategy precisely preserves the adherence of the Voronoi edges generated by the anchors to the line-segment in a consistent way. At the same time, the density and regularity of anchor distribution along a line-segment avoid influence from anchors of other line-segments.

Junction preservation. The sampling of anchors is a local procedure on individual line-segments that does not preserve their junctions. Thus *junction-anchors* are created by positioning pairs of seeds on a circle, called the *junction-circle*, centered at the intersection of the adjacent line-segments, and of radius 2ε. Anchors located inside junction-circles are first removed. Then, junction-anchors are inserted onto the junction-circle as explained in Figure 2.5. Note that three mutually adjacent line-segments are necessarily intersecting in one point as a consequence of the consolidation process. For junctions between two adjacent line-segments, the first junction-anchor is positioned at the intersection of the junction-circle with the bisector of the line-segment pair having the smaller angle. Then the other two junction-anchors are created by orthogonal symmetry with respect to the two line-segments. In term of junctions between three adjacent line-segments, the two line-segments having the smallest angle are selected to position three junction-anchors as described before. Then two more junction-anchors are symmetrically positioned to the third line segment without disturbing existing junction-anchors. In such a

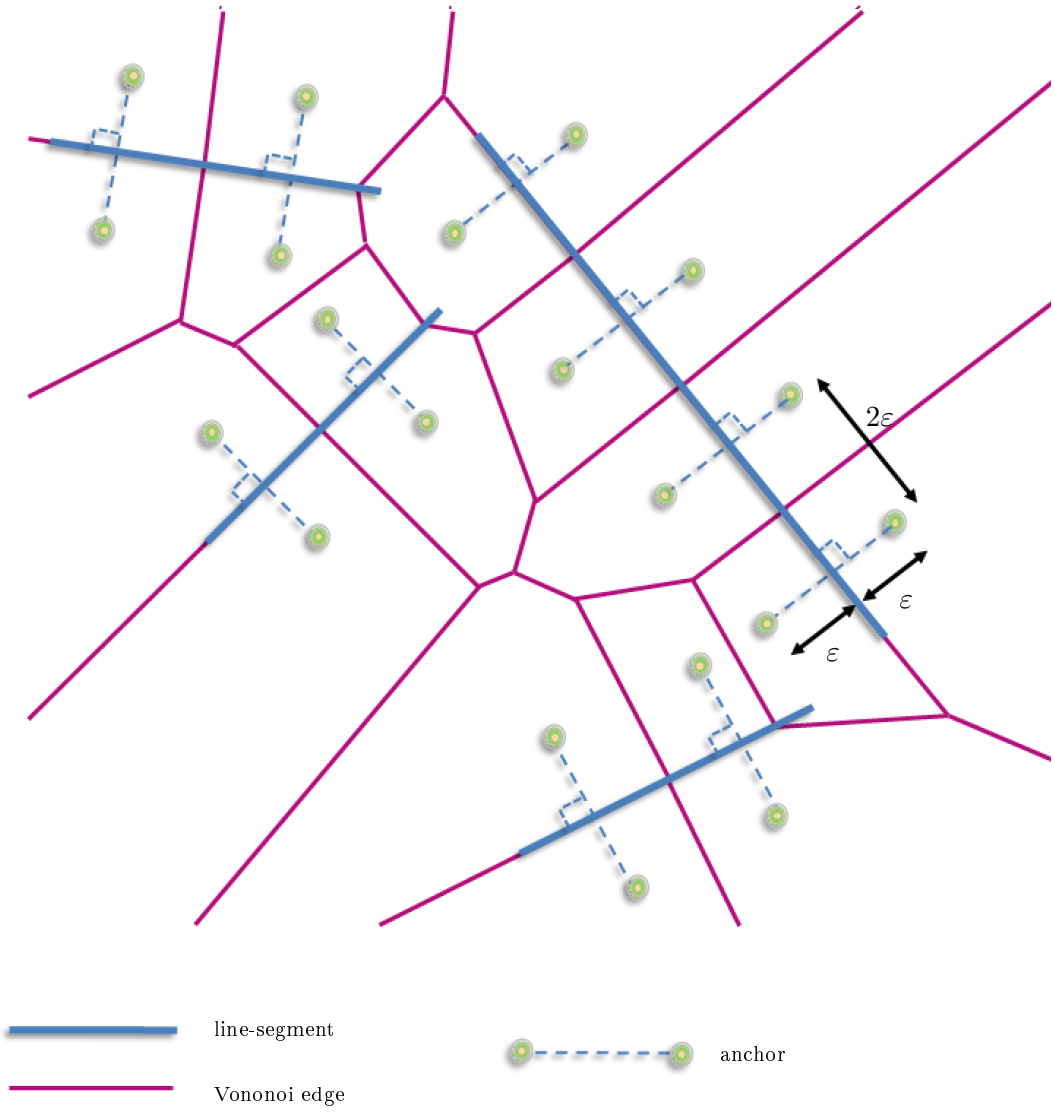


Figure 2.4: Anchoring. A set of anchors is positioned orthogonally to each line-segment, each seed of an anchor being at the same orthogonal distance ε from the line-segment. Two adjacent anchors related to the same line-segment are distant by 2ε.

way, the Voronoi edges generated from these junction-anchors exactly preserve the junction and lie on the line-segments within the junction-circle.

The anchoring procedure is entirely controlled by the parameter ε. Note that cells generated from junction-anchors have typically a triangular shape that reduce

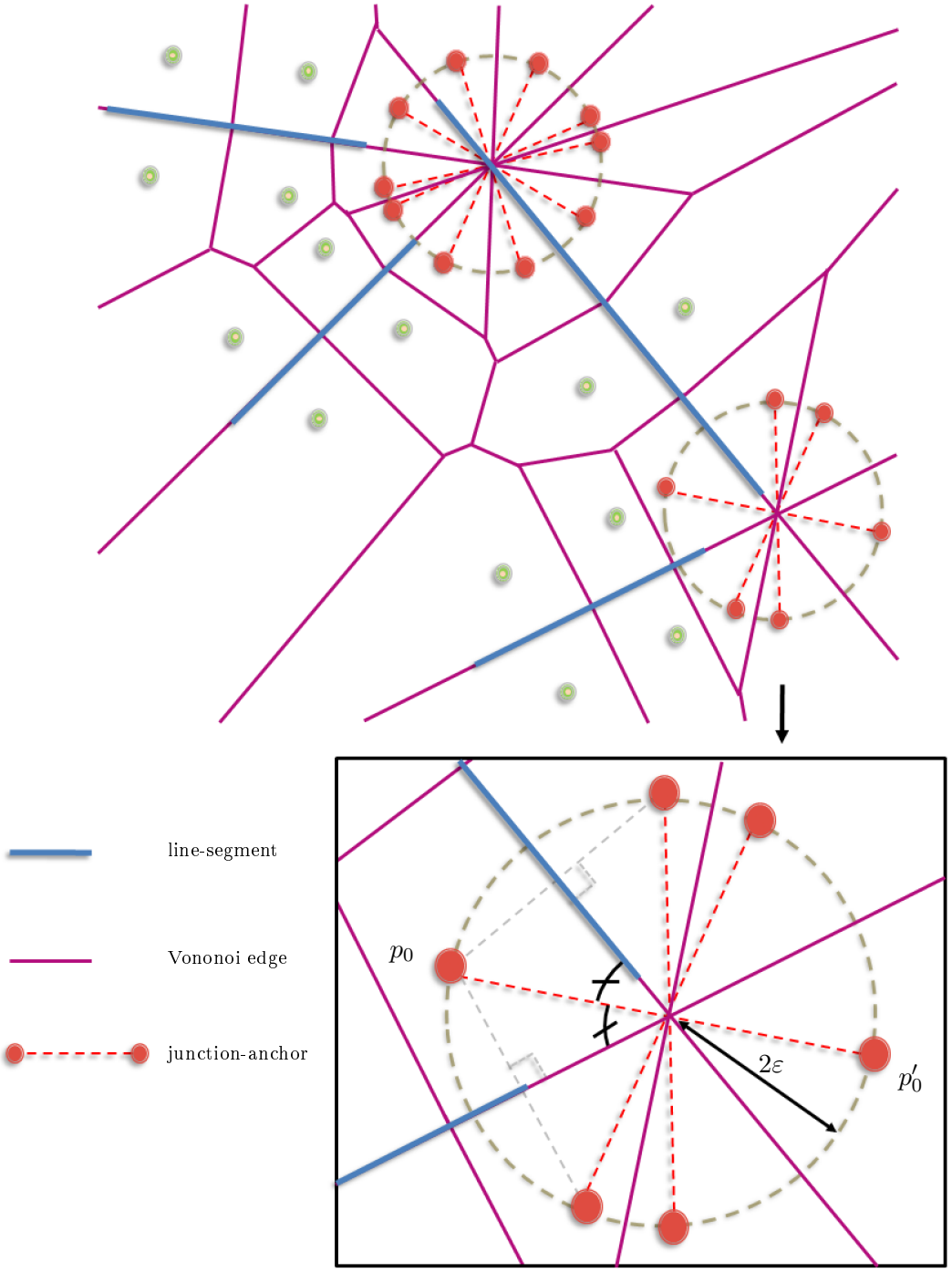


Figure 2.5: Junction-anchoring. Three (resp. five) junction-anchors are positioned to preserve junctions between two (resp. three) lines (top). The starting junction-anchor (p_0, p'_0) is positioned at the intersection of the junction-circle with the bisector of the line-segment pair having the smaller angle. Then the other junction-anchors cab be created by orthogonal symmetry with respect to the line-segments (bottom).

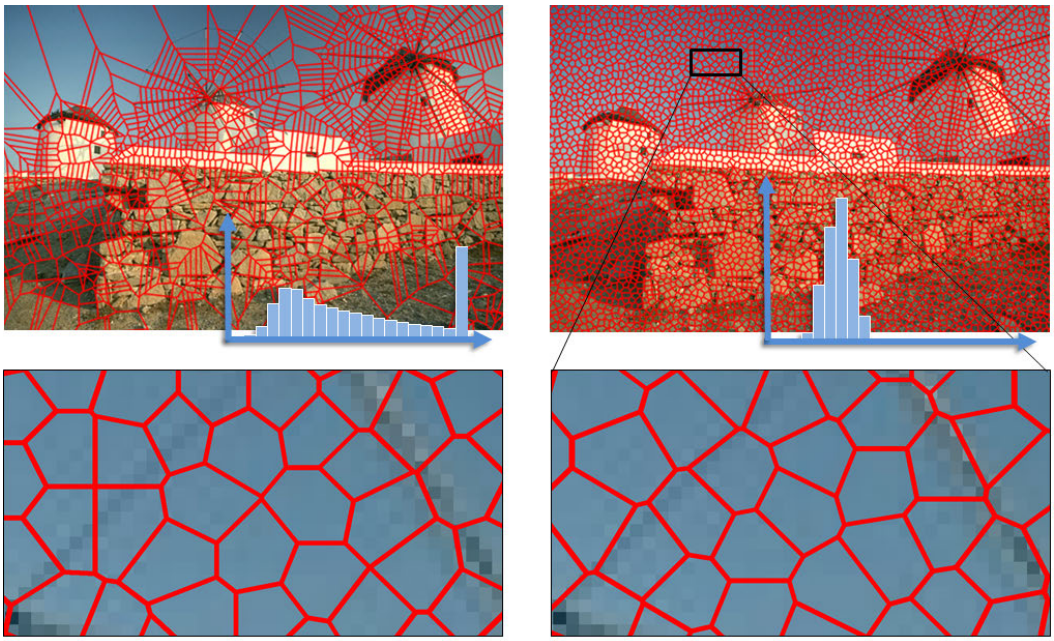
the global compactness of the partition. This is the price to pay for preserving the exact intersection of the line-segments into the partition. Junctions between at least four line-segments are marginal in practice: this case is not handled by our algorithm.

2.6 Spatial homogenization

The Voronoi partition from anchors generates cells of heterogeneous size. In particular, large cells poorly captures the homogeneous areas of the input image. If line-segments capture well the main image discontinuities, they are less adapted to secondary boundaries, as those formed by the sail frames of the windmill on Figure 2.6. Hence, the Voronoi partition is refined by sampling a point process for a better spatial homogenization of polygons.

Sampling domain. A sampling domain is first defined so that the Voronoi edges supporting line-segments and their junction will not be affected by the insertion of new seeds. This domain is defined as the complementary, over the image domain, of the accumulated disks centered on each seed and of radius 2ε.

Poisson-disk sampling. New seeds are then distributed over this domain using a Poisson-disk sampling, the disk radius being equal to ε. Instead of considering a homogeneous sampling, the seed distribution is guided with a spatial density, similarly to [BSD09]. The density is defined as proportional to the inverse of the image gradient, as detailed in Figure 2.7. The intuition behind that is to avoid new seeds to be positioned on image discontinuities. This procedure does not guarantee to produce Voronoi edges that perfectly align with secondary boundaries, but it encourages the positioning of seeds at the center of local homogeneous areas, as illustrated on Figure 2.6, bottom middle. Note that other types of spatial densities can also be used, e.g., texture or distance maps depending on the study context.



uniform sampling gradient-driven sampling

Figure 2.6: Spatial homogenization. Initial Voronoi partition from anchoring (top left) is refined into a partition (top right) with regular-sized cells (see histograms of the distribution of the Euclidean distance between boundary pixels and region centroid). When no line-segments are detected in an area, our Poisson-disk sampling driven by the image gradient allows the preservation of secondary boundaries contrary to a uniform sampling (see close-up).

2.7 Comparison with superpixel methods

The algorithm is implemented in C++, using the Computational Geometry Algorithms Library [1] for the Voronoi diagram structure as well as for the basic geometric operations as the computation of the line-segment distance. All timings are measured on an Intel Core i7 clocked at 2GHz. Experiments with size images from the Berkeley dataset are conducted to show the performance of the proposed algorithm comparing with state-of-the-art superpixel methods.

The main parameter of our algorithm, ε, allows the control of the cell size. This parameter steps in the different stages our system. Four additional parameters

1. www.cgal.org

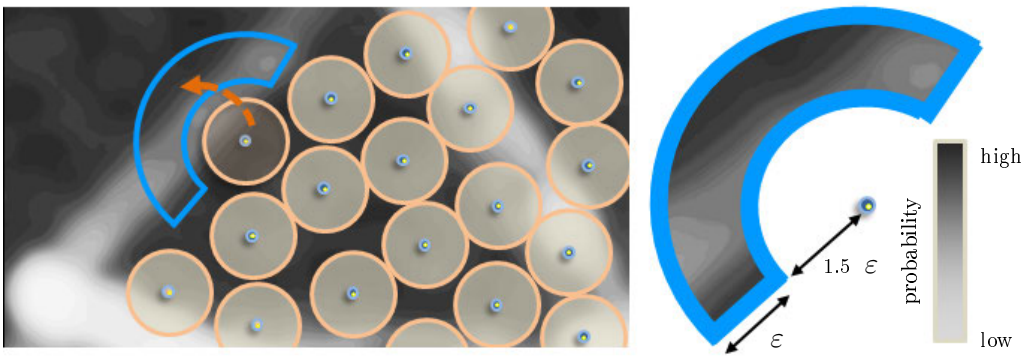


Figure 2.7: Poisson-disk sampling with non-homogeneous spatial distribution. Each new disk to insert is positioned into a circular domain of width ε (blue contour). The sampling is guided by the inverse of the image gradient (grey scale), here from the close-up of Figure 2.6.

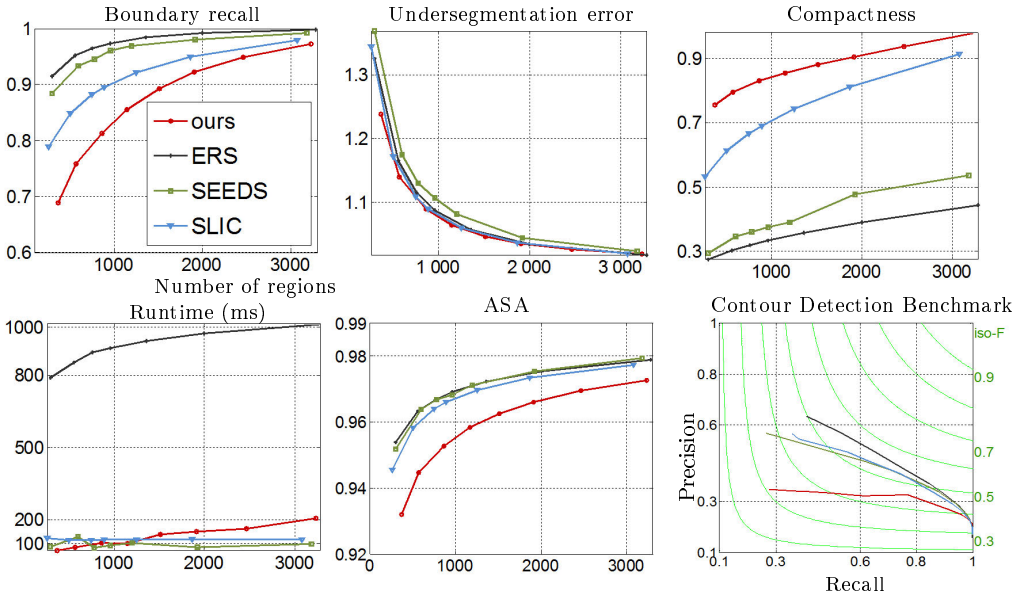
are used during line-segment consolidation (Section 3.4): a maximal angle and a maximal distance to define the near parallelism and near-colinearity of line-segments, as well as a maximal radius of inscribed circle of three line-segments, and a minimal large to small line-segment length ratio. These four parameters are fixed respectively to 5^o, 0.5ε, 0.5ε and 5 in all the experiments.

Flexibility. Because of the nature of the geometric shapes, our algorithm is particularly suitable for man-made environments in which boundaries are often accurately described by line-segments. It also produces convincing results on free-form boundary images as illustrated on Figures 2.2 and 2.9 (top row), even if the piecewise-linear approximation of object contours can be penalizing. Radiometric information are exploited at two different levels in the algorithm, i.e., during line-segment extraction and Poisson-disk sampling. The former plays a more important role as its conditions the positioning of the Voronoi edges onto the main image discontinuities.

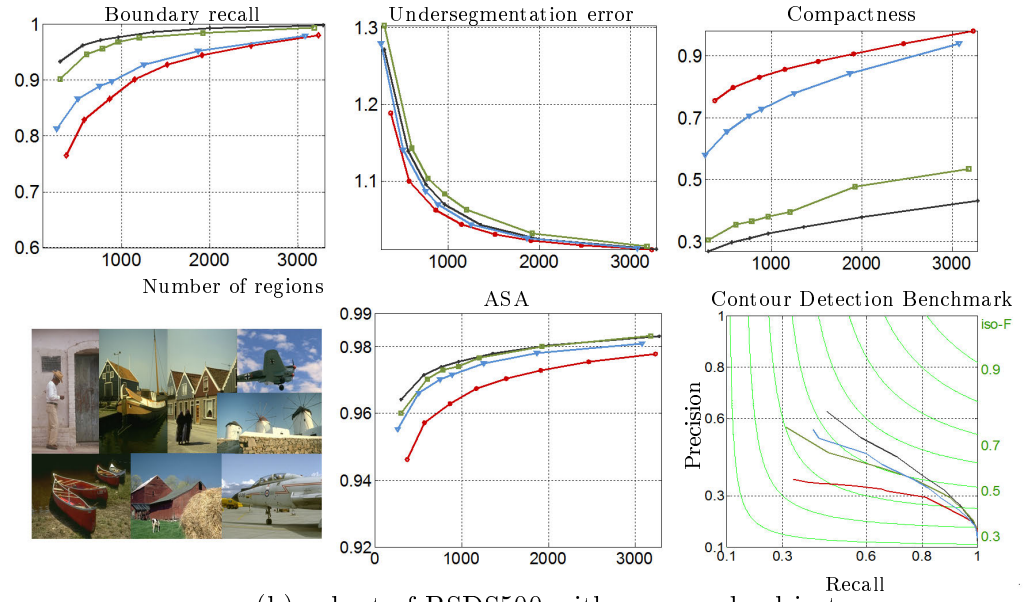
Comparison with superpixel methods. Although the proposed algorithm produces polygonal regions different from superpixels, it can be evaluated using

the standard quality criteria required for superpixel methods. Four quality criteria are taken into account: boundary recall [ASS+12], undersegmentation error [LSK+09], compactness [SFS12], and running times. Our algorithm is compared on the Berkeley dataset [MFTM01] with three state-of-the-art superpixel methods: SLIC [ASS+12], SEEDS [VdBBR+12] and ERS [LTRC11]. For measuring the quality criteria on our method, the edges of the polygonal regions are discretized into pixel-based boundaries.

Figure 2.8 shows the results on the four quality criteria. Because our regions are convex polygons of homogeneous size, our algorithm outperforms the other methods in terms of compactness by a significant margin. The algorithm also competes well in terms of undersegmentation error and running time. Contrary to SEEDS and SLIC, our running time increases in function of the number of regions. Nevertheless, as our algorithm manipulates geometric objects, it is less impacted when the image size increases. In addition, our memory consumption is very low, even on very big images. Our result on the boundary recall scores low with respect to the three other methods. The use of line-segments logically penalizes the boundary accuracy, in particular when the number of regions is low. This is the price to pay to guarantee highly compact regions, boundary recall and compactness being hard to conciliate. Nevertheless, the boundary recall of our method improves when the evaluation is restricted to a subset of images for which man-made structures are dominant, as shown in Figure 2.8, bottom row. In particular, the boundary recall becomes quite close to SLIC. For such images, the boundary accuracy is less penalized by the use of line-segments. In terms of model parameters, our algorithm does not have a weight balancing between image faithfulness and region regularity as in SLIC or ERS. On the one hand, this characteristic reduces the flexibility of our algorithm. On the other hand, it allows us to guarantee some geometric properties (polygonal shape, region convexity, unique adjacency graph) contrary to the other methods. Results from other quality criteria, the achievable segmentation accuracy (ASA) [VdBBR+12] and the precision-recall, are presented as well for a general comparison. The ASA results on the Berkeley dataset evolve in a similar way than the boundary recall tests as presented. The order of magnitude in however less important as our



(a) Berkeley BSDS500



(b) subset of BSDS500 with man-made objects

Figure 2.8: Quantitative evaluation on Berkeley dataset. Boundary recall, undersegmentation error, compactness, runtime, ASA, and precision-recall are given for the entire dataset (a), and for a subset of 30 images (b) from Berkeley dataset in which man-made structures are dominant, with some samples at the bottom left.

method is globally 1% less accurate than the best score of the three superpixel methods.

Figure 2.9 shows some visual comparisons with these three superpixel methods. Our method competes well, specially for indoor and urban scenes. If the other methods typically perform better for capturing thin irregular details with large region size, our method compensates by a higher region compactness, some geometric guarantees on the result, and region boundaries under the pixel scale as shown in Figure 2.10.

Limitations. The proposed algorithm is designed to partition images with a polygonal approximation of region boundaries. If this approximation is usually relevant for man-made environments, it might be of lower interest for images with weaker geometric signatures. The accuracy of our results is also dependent of the quality of the detected line-segments. The state-of-the-art line-segment detector [VGJMR10] is employed: it produces accurate line-segments but still lacks of global regularization to get line-segment configuration of very high quality.

2.8 Conclusion

In this chapter, a novel algorithm is proposed to partition images into convex polygons. Contrary to superpixel methods, the proposed method operates at the scale of the geometric shape and not directly at the pixel scale. Our algorithm has demonstrated several interesting properties in terms of geometric guarantees, region compactness and scalability, and has shown potential for partitioning images with strong geometric signatures, typically man-made environments. The key technical ingredient of our work is an anchoring procedure to conform Voronoi diagrams to geometric shapes, or, to be more precise, to line-segments.

This work brings a geometric dimension to traditional superpixel segmentation methods. Used as preprocessing, it is expected to serve Vision to exploit more efficiently the geometric knowledge disseminated into images, for instance by polygonalyzing objects with region grouping, classifying scenes at a subpixelic scale or

Figure 2.9: Visual comparison. The proposed algorithm produces competitive results for man-made objects or environments (four middle rows) in which the geometric structures are preserved. Only SLIC presents regions of the same order of compactness than our algorithm, but with more outliers (see the histograms representing the distribution of the Euclidean distance between boundary pixels and region centroid, for the medium region size).

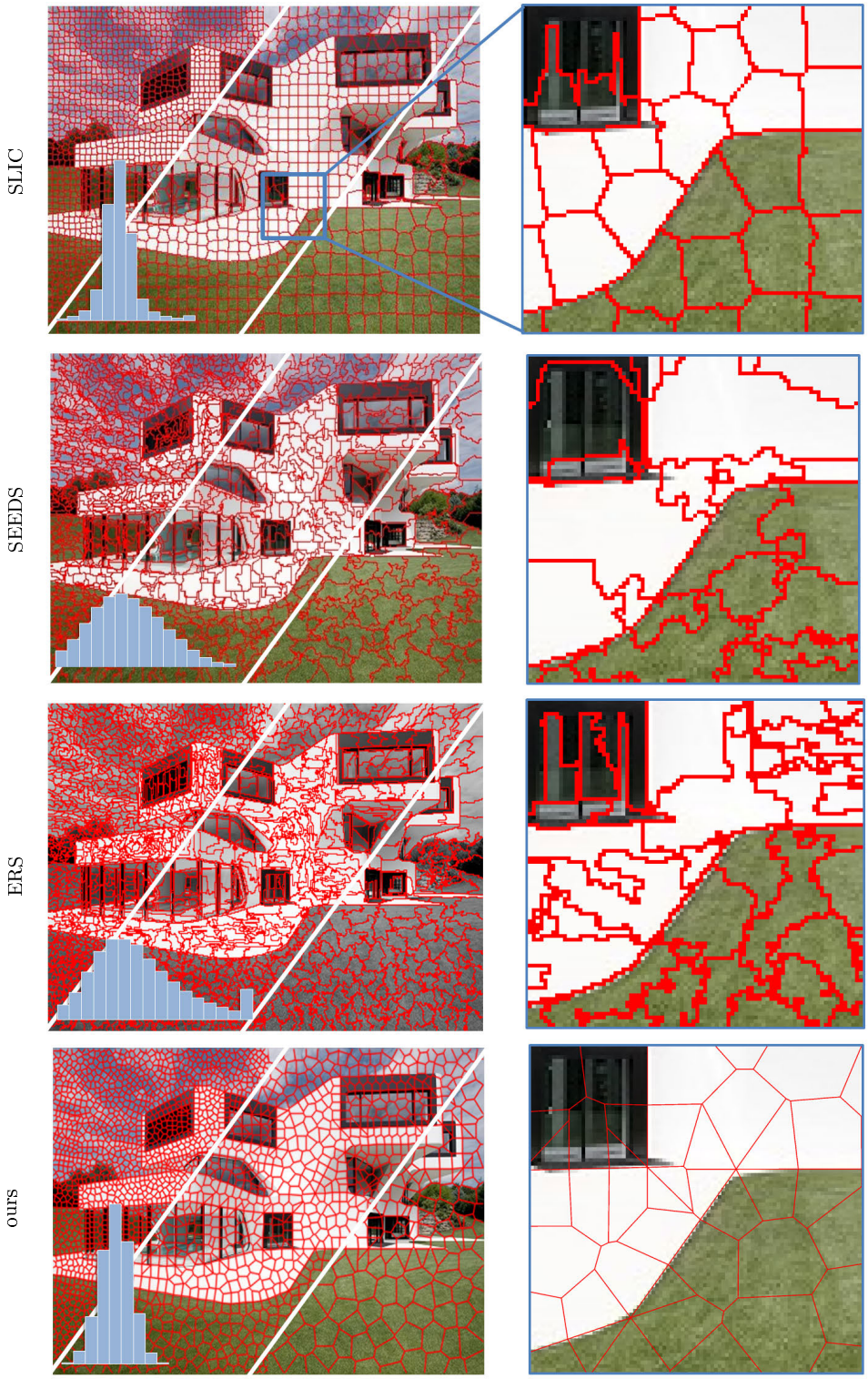


Figure 2.10: Region boundary. Contrary to traditional superpixel methods such as SLIC, SEEDS, and ERS, the regions produced by the proposed algorithm are polygons able to preserve the geometric signatures of images at a subpixelic scale shown at the closeups.

matching regions for stereo. Some applications presented in Appendix 7.1 illustrate the potential of our approach in Vision.

The use of line-segments is however not fully adapted to images with weak geometric signatures. In the future work, it is possible to investigate the use of more flexible geometric shapes that capture better free-form objects. Quadrics or B-splines are potential solutions assuming Voronoi diagram can be built in non-Euclidean space that conform to these shapes.

Chapter 3

Joint classification

Atomic regions of satellite images are generated by the polygonal partitioning algorithm demonstrated in Chapter 2, preserving geometric shapes. Inspired by region-based stereovision, a strategy is proposed for simultaneously classifying semantics and estimating elevations from polygonal partitions of a satellite stereo pair. Those polygons labeled as the semantic class *roof* are enriched with relative elevation values. Figure 3.1 illustrates the goal.

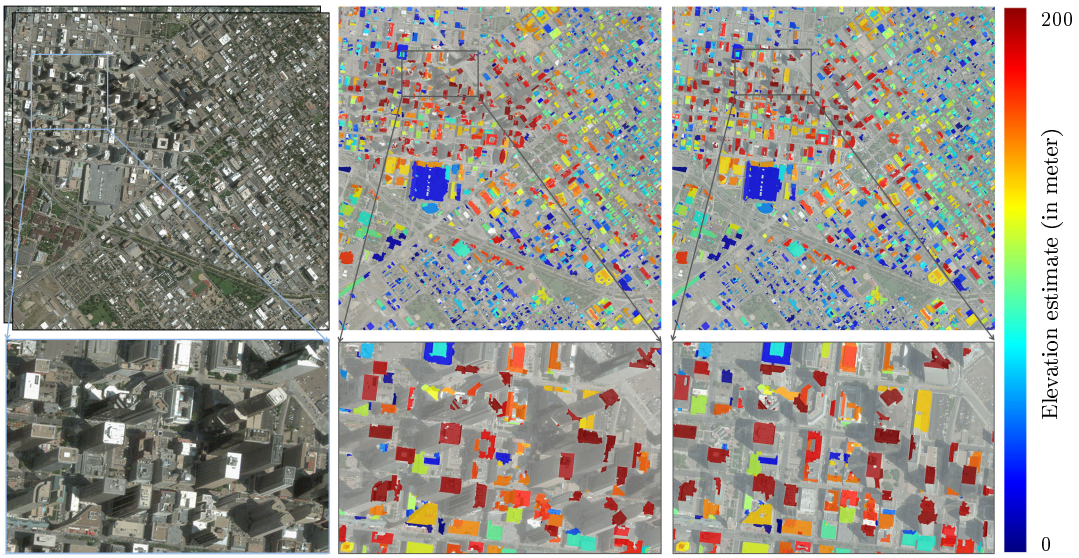


Figure 3.1: Joint classification of Denver downtown. Starting from a stereo pair of satellite images (left), our algorithm produces a semantic classification with elevation simultaneously estimated (marked with gradient color, right).

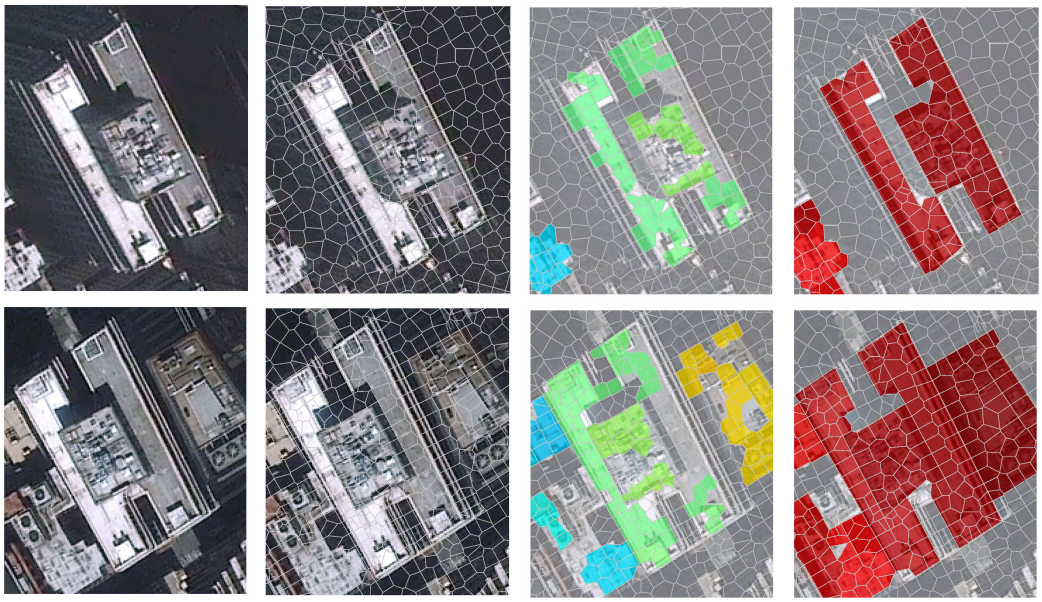
3.1 Introduction

Image understanding based on image semantics has been an important research topic in Computer Vision for many years. Semantic classification plays a valuable role in retrieval of meaningful information in images, especially in urban scenes. Traditional classification approaches mostly rely on the matching of visual features, such as texture, intensity and color information, to classify pixels into semantic classes [HTP05, PTN09, KOSPK16]. Contextual classification methods integrate spatial consistency and neighboring information to better capture object-level targets in high resolution imagery, while suffering from complex feature design and the size of context regions [Mni13]. Moreover, intra-class variations can be large in satellite city scenes, such that the contrast among parts of a roof can be higher than the contrast between the roof and its surroundings [AA10]. Only relying on 2D appearance clues is not sufficiently effective to classify urban context from satellite imagery. Another useful potential clue is height information providing a 3D interpretation of scenes [KMRB09]. However, stereo matching of satellite images produces sparse elevation maps that are incomplete and rough for semantic extraction as explained in Section 1.4.

A joint strategy is proposed to retrieve semantics and estimate elevations simultaneously, by interacting radiometric and 3D geometric information both from the left and the right images. The output consists of polygons labeled with semantic classes: *roof* and *other*, and enriched with an elevation value for each polygon determined as *roof*. The overview is illustrated on Figure 3.2.

3.2 Review of region-based stereo matching

Numerous works have been proposed in stereo matching [SS02]. While well-established methods as the SGM algorithm [Hir08] reason at the scale of pixel, some works focus on matching image regions to more accurately preserve object boundaries [ZK07a, TWZ08]. Beyond boundary accuracy, region-based stereo matching methods can offer high scalability and time-efficiency [BSRG14]. Some works



(a) (b) (c) (d)

Figure 3.2: Overview of the joint classification. Input stereo images (a) are first decomposed into atomic convex polygons (b) using the algorithm described in Chapter 2. Secondly, the initial elevation of each polygon is calculated by the SGM [Hir08] with double-checking (c). In the last step detailed in Section 3.4, the semantic class and the elevation of each polygon are simultaneously retrieved in the two sets of partitions (d). The color reflects the corresponding elevation value in a proportion to the elevation range of the scene, which is different in (c) and (d) since elevations are quantized in (d).

[BRK+11, LSR+12] also combine object segmentation or classification with stereo matching in unified frameworks. Inference for these models is, however, a complex task that requires time-consuming optimization procedures. Overall, most of these methods are not adapted to satellite images whose wide baselines typically produce severe occlusion problems that are not specifically handled. The additional use of geometric primitives as line-segments usually helps to better interpret occluded parts of images [BFVG05].

3.3 Elevation assignment to polygonal partitions

The algorithm in Chapter 2 is applied independently on both satellite stereo images with a polygon size fixed to 5 pixels, i.e., average distance of polygon edges to its polygon centroid) in our experiments. As illustrated on Figure 3.3, it captures geometric regularities in images by aligning contours of atomic polygons with linear structures such as roof edges. Note that the line-segments embedded into the polygonal partitions will be used further in our approach.

The polygons are enriched with an *elevation estimate* which corresponds to the altimetric distance between the observed surface captured in the polygon and the ground. For each polygon, its elevation estimate is defined as the difference between the mean of the pixel depths contained inside the polygon (computed by SGM [Hir08] with double checking), and the depth of the ground (computed by a standard Digital Terrain Model (DTM) estimation method [BPD02]). Because of the wide baseline of our stereo pairs, polygons without elevation estimates are frequent, especially when associated to facade elements as illustrated in Figure 3.3. In return, elevation estimates are relatively accurate and present on a very large majority of roofs. Our strategy is thus to couple these elevation estimates with the geometric information contained in the polygonal partitions to retrieve building contours even for partially occluded roofs.

The polygonal partitions produced by the algorithm in Chapter 2 are denoted by $\mathcal{P}_l$ and $\mathcal{P}_r$ for the left and the right images respectively. $\mathcal{P}_l^\star \subset \mathcal{P}_l$ represents the set of polygons in $\mathcal{P}_l$ with elevation estimates. A polygon $i \in \mathcal{P}_l \cup \mathcal{P}_r$ associated with an elevation estimate d_i is projected in 3D using the traditional Rational Polynomial Coefficients (RPC) model [HS97]. Two polygons $i \in \mathcal{P}_l$ and $j \in \mathcal{P}_r$ with respective elevation estimates d_i and d_j are said to be *imbricate* if the orthographic projections into the horizontal plane of the 3D polygons overlap. In this case, $\tau_{ij} \in [0, 1]$ denotes the overlapping ratio of the orthographic projections, i.e., the intersection area to union area ratio. These notations are illustrated in Figure 3.4.

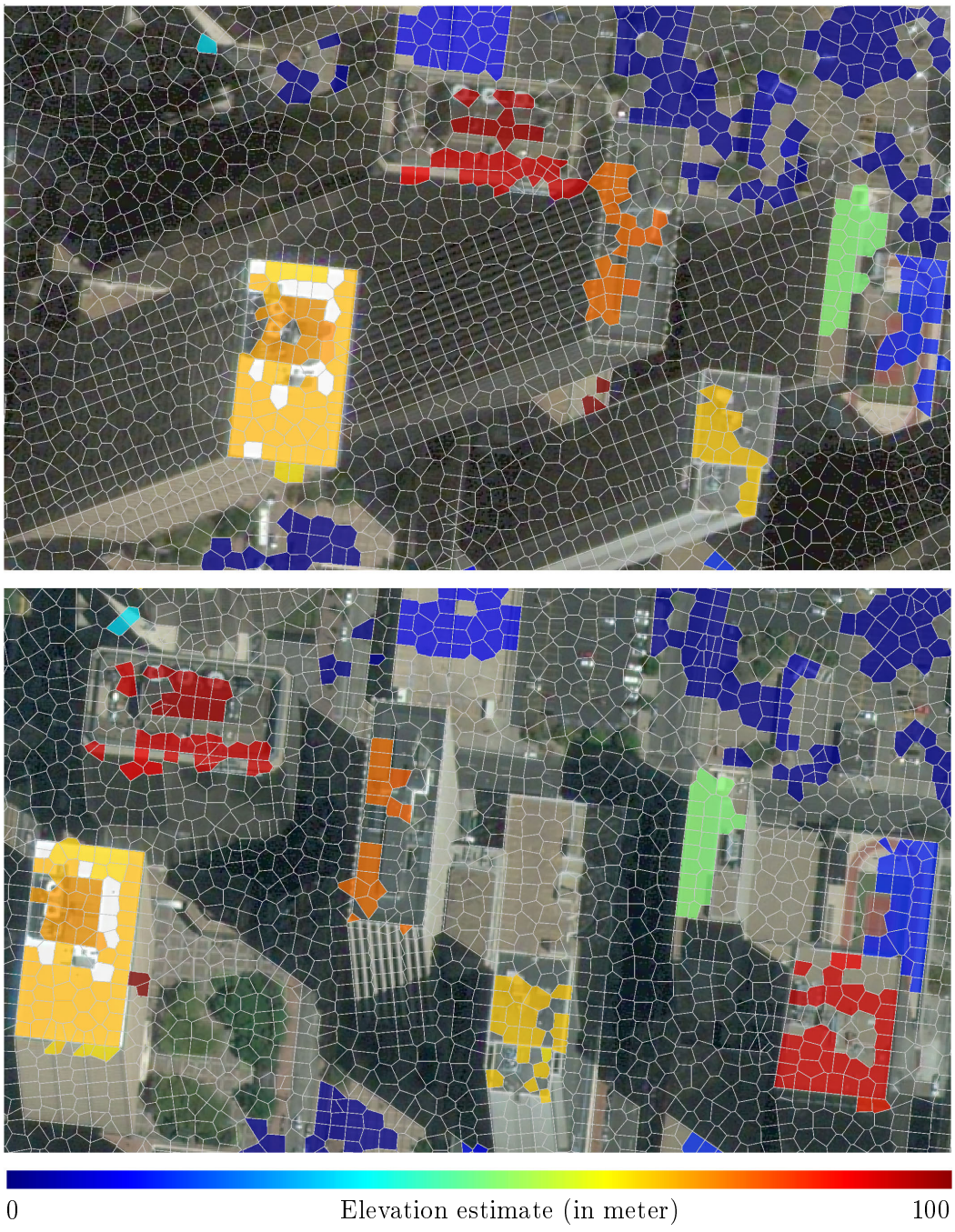


Figure 3.3: Polygonal partitioning and elevation estimates. Left and right polygonal partitions capture linear structures contained in input images, and in particular building edges. Elevation estimates sparsely cover the polygonal partitions (see colored polygons). Each roof contains at least a few elevation estimates.

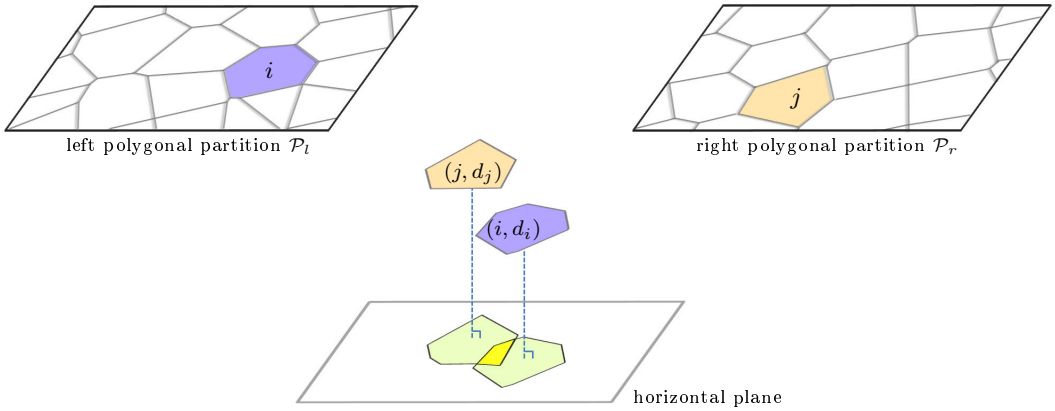


Figure 3.4: Orthographic projection of polygons. Polygons i and j with respective elevation estimates d_i and d_j are projected in 3D using the RPC model. These two polygons are imbricate as their orthographic projections into the horizontal plane overlap (see yellow area).

3.4 Joint classification and elevation recovery

Starting from the two polygonal partitions and sparsely distributed elevation estimates, our goal is now to retrieve simultaneously the semantic class and the elevation of each polygon of the partitions.

Two semantic classes of interest are considered: *roof* and *other*. Class *other* mainly refers to ground and facade elements. Because of the wide baseline, most of these elements are only visible in one image. As our main objective is to reconstruct buildings, considering only these two classes is sufficient under the assumption that facades are vertical. Contrary to class *other*, class *roof* is associated with an elevation value. By considering the classification problem as a labeling formulation, the set of possible labels can then be defined as $L = \{z_1, ..., z_n, other\}$ where $z_1, ..., z_n$ are the n possible elevation values of a roof. To set $z_1, ..., z_n$, the set of elevation estimates are clustered by K-means with $K = n + 1$, and associate the n highest centroids to them. As the ground falls into the class *other*, the centroid with the lowest value is reset to zero. A $\sigma(z_k)$ denotes the standard deviation of the k^{th} cluster.

The quality of a configuration of labels $l \in L^{card(\mathcal{P})}$ is measured through an

energy U of the form:

$$U(l) = \sum_{i \in \mathcal{P}} D_{data}(l_i) + \beta_1 \sum_{(i,j) \in \mathcal{E}_s} V_{smoothness}(l_i, l_j) + \beta_2 \sum_{(i,j) \in \mathcal{E}_c} V_{coupling}(l_i, l_j) \quad (3.1)$$

where D_{data} is the unary data term, and $V_{smoothness}$ and $V_{coupling}$ are pairwise potentials favoring respectively label smoothness and label coherence between left and right partitions. $\mathcal{E}_s$ and $\mathcal{E}_c$ correspond to two sets of pairs of adjacent polygons. β_1 and β_2 are parameters weighting the three terms of the energy.

Polygon adjacency. The two adjacency sets $\mathcal{E}_s$ and $\mathcal{E}_c$ impose spatial dependencies between polygons, either within the same polygonal partition for the former or in between the polygonal partitions for the later, as illustrated on Figure 3.5.
$\mathcal{E}_s$ contains pairs of polygons who share a common edge which is not supported by one of the line-segments embedded into the polygonal partitions. As illustrated in Figure 3.5 (right), this condition on line-segments is particularly efficient for stopping label propagation when meeting building edges.
$\mathcal{E}_c$ is defined as the set of imbricate polygons, i.e., the pairs of polygons $i \in \mathcal{P}_l^\star$ and $j \in \mathcal{P}_r^\star$ so that $\tau_{ij} > 0$.

Data term. It measures the coherence between the elevation estimate of a polygon and its proposed label. For polygons without an elevation estimate, the occurrence of the label *other* is favored as a polygon without a depth estimate is most likely to capture an element visible only in one image such as facade and, to a lesser extent, ground. The data term is expressed as

$$D_{data}(l_i) = \begin{cases} 1 - e^{\frac{-(l_i - d_i)^2}{2\sigma(l_i)^2}} & \text{if } i \in \mathcal{P}^\star \\ \alpha \cdot \mathbb{1}_{\{l_i \neq other\}} & \text{otherwise} \end{cases} \quad (3.2)$$

where d_i is the depth estimate of polygon i, $\mathbb{1}_{\{.\}}$ is the characteristic function, and α is the penalty weight for not choosing *other*. When label *other* is attributed to polygon $i \in \mathcal{P}^\star$, l_i is set to 0.

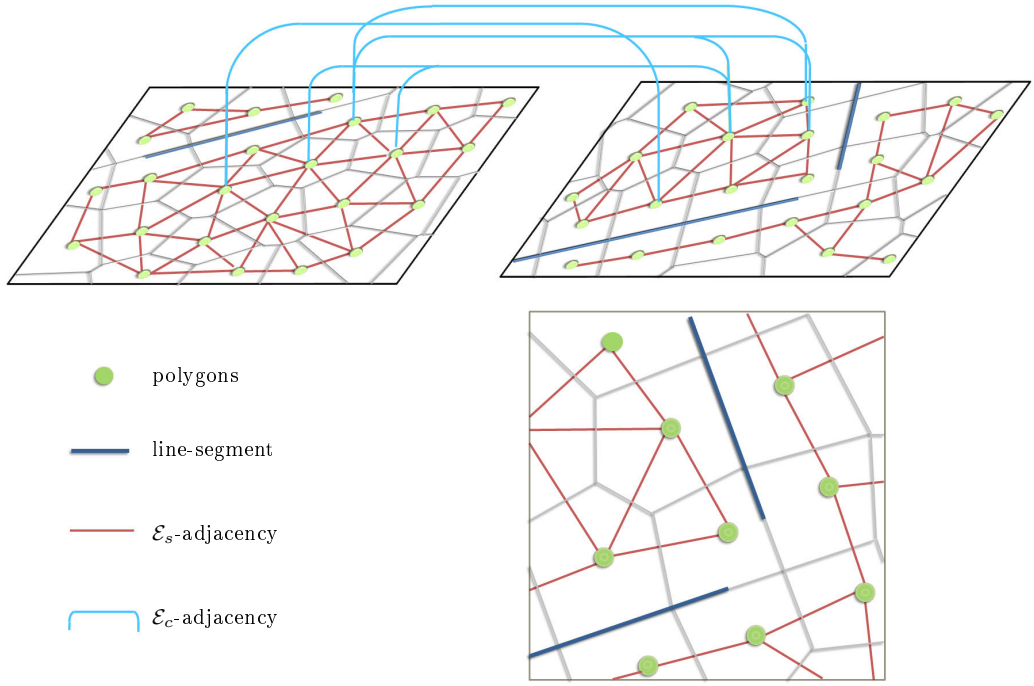


Figure 3.5: Polygon adjacency. Two types of pairwise interactions between polygons are taken into account in the labeling formulation: within the same partition and in between partitions (left). Line-segments embedded into the partitions prevent neighboring polygons from interacting (right).

Smoothness. The smoothness term penalizes $\mathcal{E}_s$-adjacent polygons with different labels using a generalized Potts model:

$$V_{smoothness}(l_i, l_j) = w_{ij} \cdot \mathbb{1}_{\{l_i \neq l_j\}} \tag{3.3}$$

$$w_{ij} = 1.0 - d_{L^2}(F_{rgbhistogram}(i), F_{rgbhistogram}(j)) \tag{3.4}$$

where the weight w_{ij} reduces the penalty of having different labels when the radiometry of pixels inside the two polygons is not similar. In practice, w_{ij} is chosen as one minus the normalized histogram distance in norm L_2. Different basic features including the average, median of LAB colors, of RGB colors, and the histogram of RGB colors can be used to measure the similarity of the radiometry. The histogram of RGB gives the most robust performance in most of our experiments in terms of

the accuracy of discriminating rooftops from the others, especially from facades that have similar textures with roofs.

Coupling. Similarly to the smoothness potential, the coupling term is defined by a generalized Potts model, here, between imbricate polygons.

$$V_{coupling}(l_i, l_j) = \tau_{ij} \cdot \mathbb{1}_{\{l_i \neq l_j\}} \tag{3.5}$$

where τ_{ij} allows polygons with different labels to be penalized proportionally to their overlapping ratio.

Optimization. An approximation of the global minimum of the energy is found using the α-β swap algorithm [BK04]. Figure 3.6 shows the impact of the different terms of the energy. In the sequel, the term *enriched* partition is defined to represent a polygonal partition whose polygons have received a class and eventually an elevation value by this energy minimization.

Joint classification of urban cities. The proposed joint classification algorithm has been applied to several cities with different landscapes. Figure 3.7 shows classification results of dense downtown in Seoul, South Korea, US downtown in New York City, the US, and antique city in Alexandria, Egypt. Buildings are globally well detected with estimated elevation values. Antique cities such as Alexandria are challenging for the narrow streets that bring difficulties in separating them in classification. Note that, semantics and elevations are simultaneously retrieved in the classification, allowing for robustly handling occlusion areas (see crops in New York City in Figure 3.7).

3.5 Conclusion

This chapter proposes a joint classification method for semantic object labeling and elevation estimating from satellite stereo images. Semantics and elevation estimates are jointly retrieved to be robust for handling low image quality, limited resolution and occlusion areas in satellite images. Semantics in large-scale urban

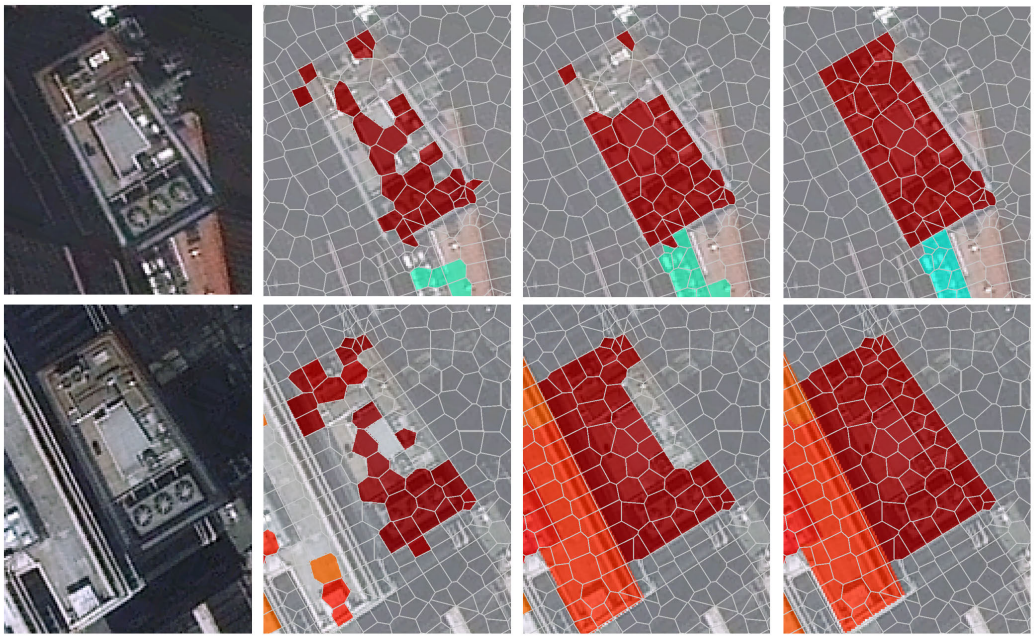


Figure 3.6: Impact of the different energy terms. Roofs are sparsely labeled using the data term only ($\beta_1 = \beta_2 = 0$, 2^{nd} column). Adding the smoothness potential propagates roof labels while preserving building edges ($\beta_2 = 0$, 3^{rd} column). The labeling coherence between the left and right partitions is enforced considering the complete energy formulation (4^{th} column).

scenes are effectively classified by integrating elevation information. Rooftops are extracted and simultaneously enriched with elevation values. The integration of semantics and elevations, and the interaction of information from the left and the right images, improve the robustness to the low SNR and the occlusion problems in satellite imagery. Two preliminary 3D models in LOD1 can be generated by projecting the enriched polygons in 3D using the RPC models.

Note that a learned radiometric prior to distinguish roofs, roads, vegetation and *others*, has been evaluated for the joint classification. However, the results show that local radiometric features of polygonal regions are not sufficiently reliable for robust classification of urban scenes, especially when facades and roads have very similar appearances with roofs.

For the future work, more semantic classes can be included into the classification,

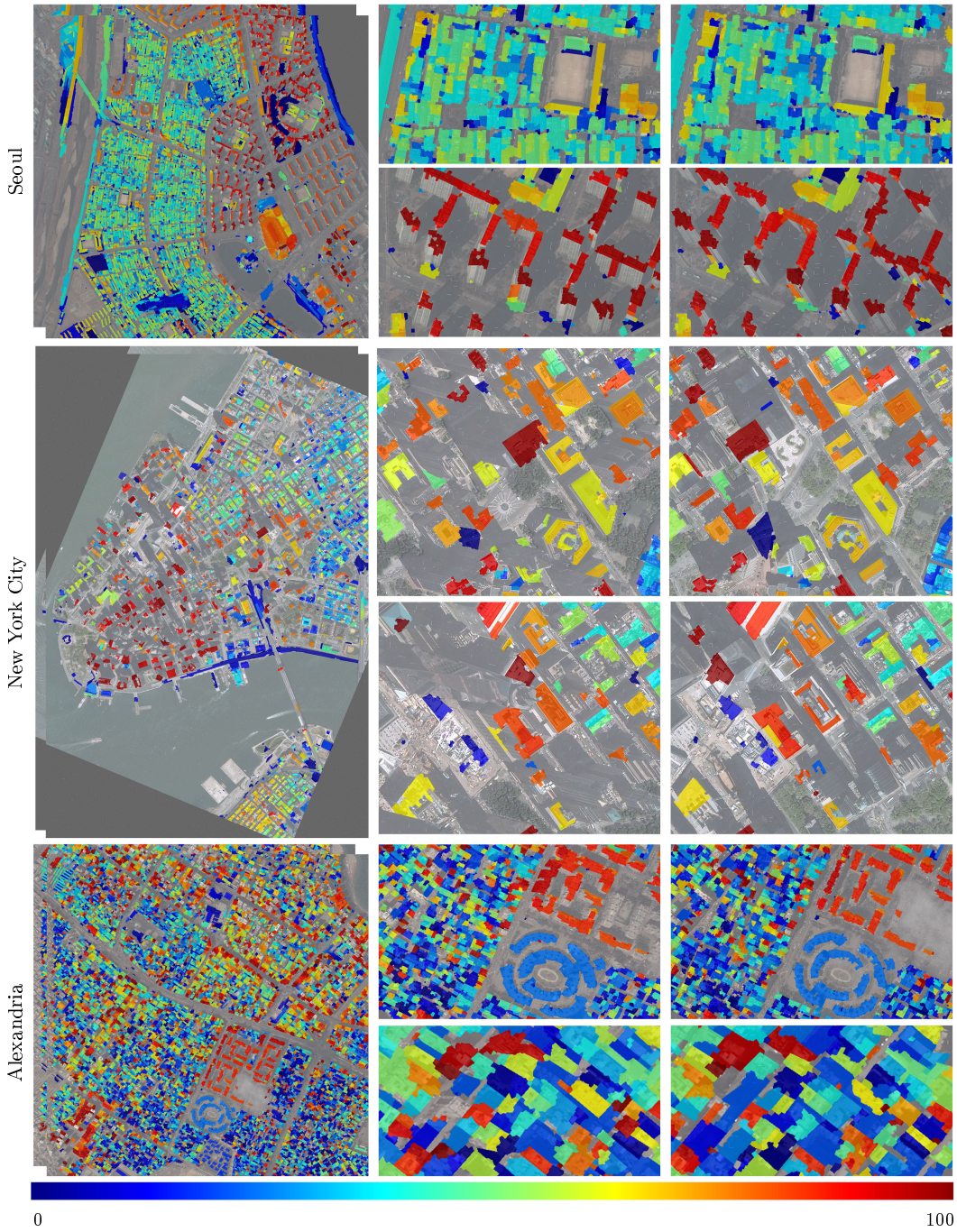


Figure 3.7: Joint classification of different cities. From the left to the right columns are: the classification of a city with elevation estimated on rooftops, and crops of the classification respectively from the left and the right images. The joint classification algorithm is performed on worldwide cities in different urban landscapes, including dense downtown with close residential buildings (top), US downtown with large, tall buildings and occlusions (middle), and antique city with narrow streets and small buildings (bottom).

in particular roads and vegetation. Learning priors with nonlocal features, such as pair-wise feature descriptors involving neighborhood information, show potential to improve the classification performance.

Chapter 4

Model fusion

Two sets of enriched partitions are produced by the joint classification algorithm proposed in Chapter 3, respectively from the left and the right images in a satellite stereo pair. Projecting these enriched partitions into 3D by the RPC models results in two different models, because of different interpretations of the shapes of objects:

- Some roof parts are frequently occluded between the two images.
- The shapes of polygons between left and right partitions do not necessarily correspond.
- The smooth term of Equation 3.1 depends on the radiometric appearance which is different in each individual image.
- The coupling term of Equation 3.1 is a soft constraint that does not guarantee that imbricate polygons have the same elevation.

Figure 4.1 shows an example of the two different 3D interpretations of a building.

In this chapter, a model fusion method is proposed to produce compact and geometrically accurate 3D city models from the two preliminary models generated by the joint classification algorithm. It relabels the polygons by integrating constraints of radiometric discontinuities and geometric shape regularities.

4.1 Introduction

Geometric accuracy of 3D models is an important evaluation metric for high quality city reconstruction. The quality of the two preliminary models from the joint classification is not sufficient for applications that require high compactness and geometric accuracy. Each of the models may capture parts of the geometric shapes well, but neither gives precise or complete object contours.

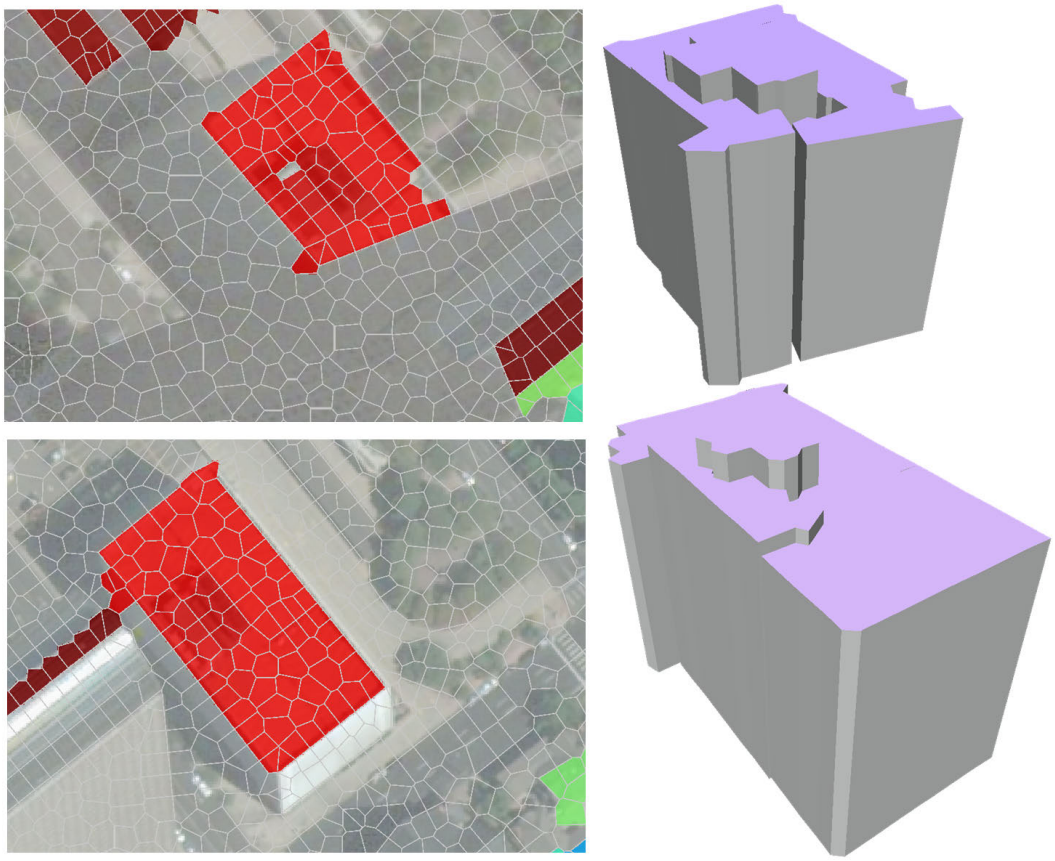


Figure 4.1: Models from the joint classification of the left and right images. The projection in 3D of the left and the right enriched partitions is not necessarily coherent with each other in terms of object shapes. Top row shows the interpretation produced by the joint classification algorithm from the left image, in which the building in the middle is partially occluded. This 3D model captures well the boundary of the higher roof layer in the middle, but is incomplete due to the occlusion area. Bottom row shows the results from the right image, in which the 3D model gives a complete contour of the rooftop, but misses the top left corner.

In a LOD1 interpretation, roofs are represented by flat horizontal primitives, and contours in the footprint indicate the borders of different height layers. Hence, a high quality LOD1 model reflects *accurate* object contours in the orthogonal projection since facades are vertical.

To unify such two preliminary models into a unique geometrically accurate 3D

representation, all enriched polygons are projected into the horizontal plane, and relabel elevations inside the new induced horizontal partition. This method retrieves geometry and semantics simultaneously, improving geometric accuracy, and robustness to occlusion areas, low image quality, and model conflicts. Figure 4.2 illustrates the fusion process from the joint classification of the two input stereo images (center).

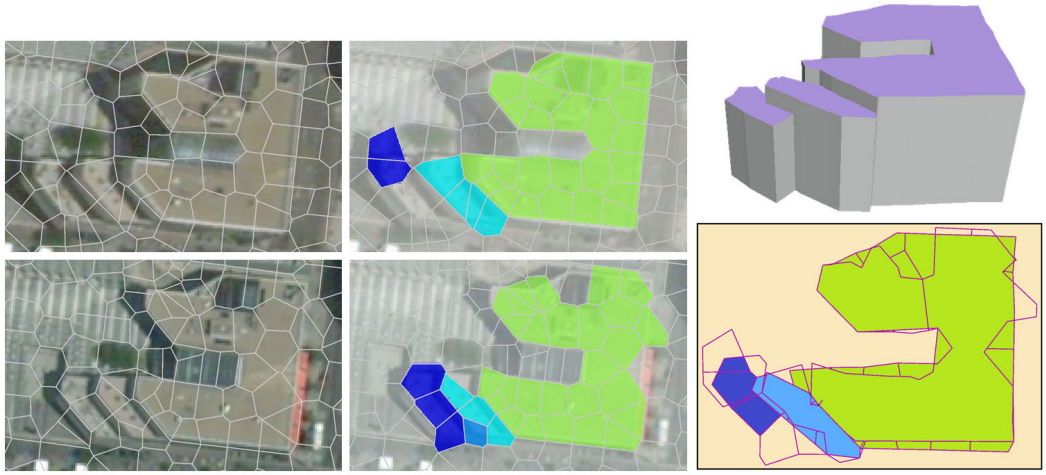


Figure 4.2: Overview of model fusion. Based on the joint classification of the input stereo pair, polygons are enriched with semantic labels and elevations values, called rough models (2^{nd} column). The fusion of these two models operates on their projections to a horizontal plane (3^{rd} column bottom), and generates the compact and geometric accurate 3D model (3^{rd} column top).

4.2 Review of object contouring

The review of previous work covers two main facets of the model fusion problem: object polygonalization and contour optimization.

Object polygonalization. Capturing objects by polygonal shapes provides a compact and structure-aware representation of the object contours. It is particularly adapted to representing regular objects as roofs from images. Object polygonalization methods typically depart from the detection of line-segments which are then

assembled into polygons. This second step can be done, for instance, by searching for cycles in a graph of line-segments [ZFW+12], or by connecting line-segments with a gap filling strategy [SCF14]. Grouping atomic regions [LSD10] is also a possible approach, especially when the number of objects is high, and the input image is big. It requires, however, a post-processing step to vectorize chains of pixels into polygons with typically a loss of accuracy.

Path optimization. Searching for optimal contours/paths provides a technique for formulating man-made object shapes. In urban cities, footprints of objects, such as buildings and roads, mostly consist of regular geometric primitives including straight lines and curves. Given a graph whose edges indicate possible paths to construct object contours, a path optimization problem can be formulated to explore the optimal contour. Works such as [CF14] reconstruct regular shapes from a free-space evidence by solving a shortest path problem through dynamic programming instead of Dijkstra's algorithm. Priors can also be introduced to bias the path searching as described in [WLH06], exploring optimal paths to formulate footprints of buildings in a space of boundary points. Most of the works are focused on extracting outlines of objects, but have less considerations on internal structures, for instance, elements constructed inside the outline of roofs.

4.3 Fusion of enriched partitions

For city reconstruction in LOD1, accurate shapes of rooftops are the most critical targets. The proposed model fusion method optimizes object contours on the orthogonal projection of the enriched polygons in 2D. The edges of adjacent polygonal projections who have different elevation values form the contours of rooftops, including outlines and shapes of their interior height layers.

Orthographic projection. Each polygon $i \in \mathcal{P}$ whose class is not *other* is projected into the horizontal plane. The superposition of projected polygons from left and right partitions produces a decomposition of the horizontal plane into new polygons called *cells*. Note that the cells are not necessarily convex. The set of cells is

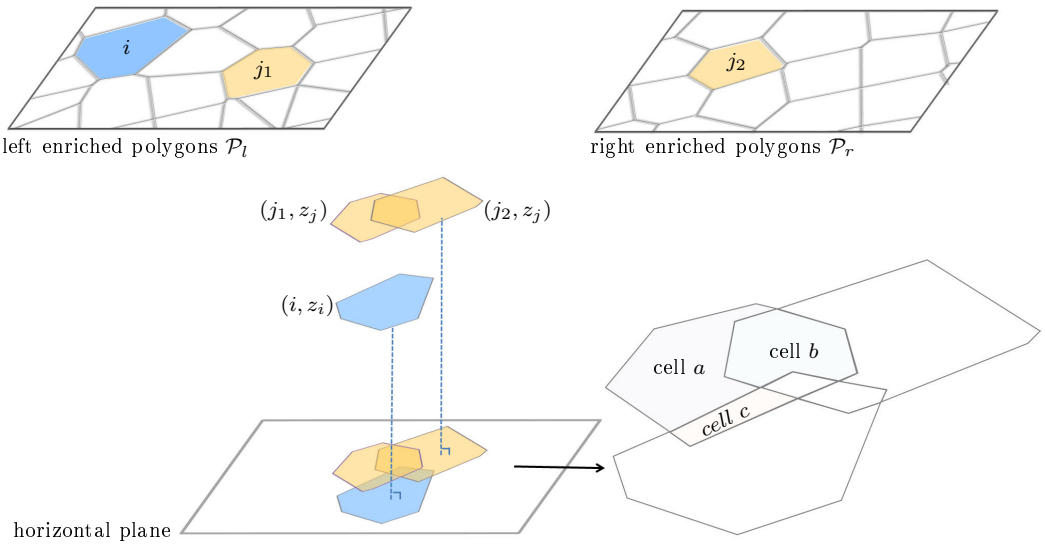


Figure 4.3: Elevation inherency. Enriched polygons i, j_1, and j_2 respectively with elevation estimates z_i, z_j, and z_j are projected into 3D using the RPC models, and then are orthogonally projected into a horizontal plane. Each cell inherits a set of elevations from the polygons that overlap with it. Cell a shares no overlap with other polygons, such that $Z_a = \{z_j\}$. Polygons j_1, j_2 overlap at cell b with identical elevation value z_j, so $Z_b = \{z_j\}$. For cell c, polygons i, j_1 with two different elevations z_i and z_j overlapped with it, hence $Z_c = \{z_i, z_j\}$.

denoted as $\mathcal{C}$. Each cell inherits the elevations of the polygons that overlap with it. The set of elevations inherited by cell $i \in \mathcal{C}$, is denoted as Z_k. Figure 4.3 explains the process of elevation inherency. Different types of cells can be distinguished:

- **Coherent cells** are cells that inherit two identical elevations, one from the left partition and one from the right. The elevation value of these cells is not modified further.
- **Conflict cells** are cells that inherit at least one elevation, and that are not coherent cells.
- **Empty cells** are cells without inherited elevation. Theses cells, which typically fill in the holes in the cell decomposition, mainly corresponds to ground or small roof parts.

These three sets of cells are respectively denoted by $\mathcal{C}_{coherent}$, $\mathcal{C}_{conflict}$ and $\mathcal{C}_{empty}$, illustrated in Figure 4.4.

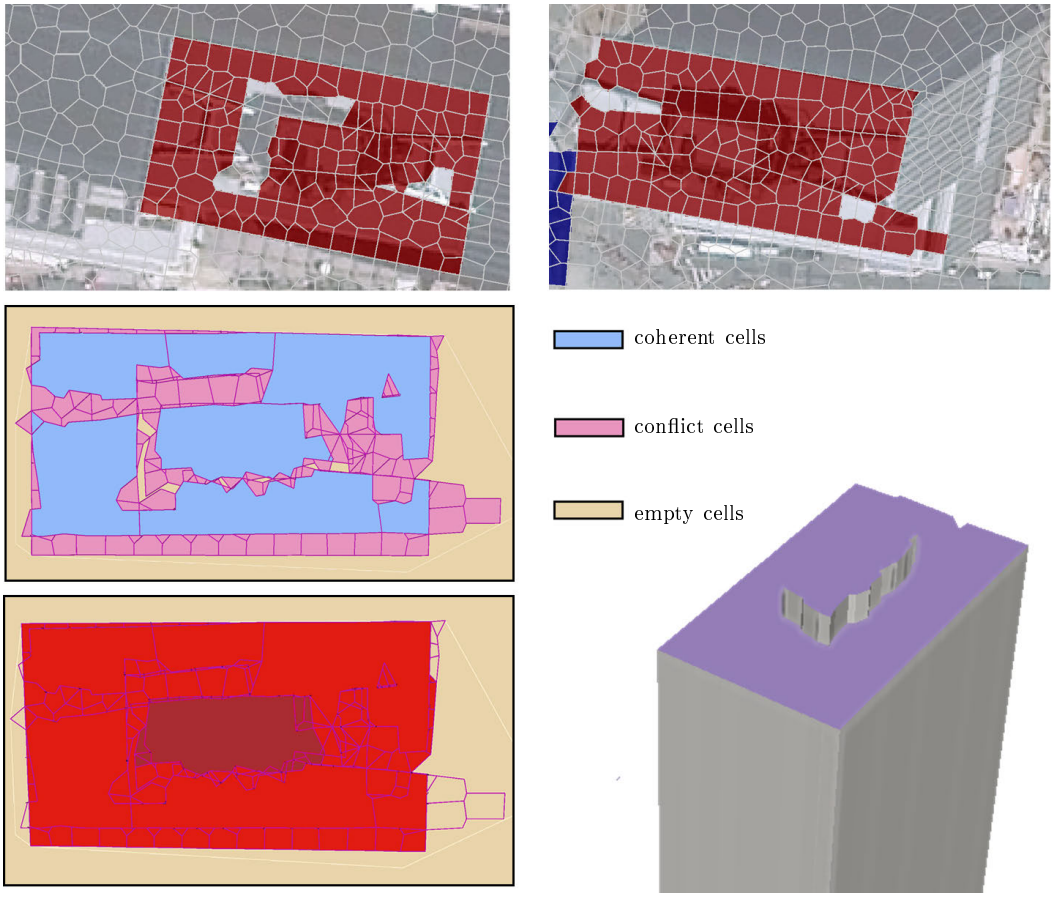


Figure 4.4: Fusion of enriched partitions. Projecting the enriched partitions (top) into the horizontal plane produces a cell decomposition in which three groups of cells can be distinguished (middle). The relabeling of the elevation of conflict and empty cells gives a unified 3D model (right).

Cell relabeling. For fusing enriched partitions, each conflict or empty cell must be associated with a unique elevation. Those cells are relabeled using an energy formulation with a standard form:

$$E(x) = \sum_{k \in \mathcal{C}^\star} A_k \cdot E_d(x_k) + \lambda \sum_{(k,k') \in \mathcal{N}} L_{kk'} \cdot E_r(x_k, x_{k'}) \tag{4.1}$$

where $\mathcal{C}^\star = \mathcal{C}_{conflict} \cup \mathcal{C}_{empty}$, the label x_k of cell k is an elevation value in $Z = \{0, z_1, ...z_n\}$, and $\mathcal{N}$ is the set of pairs of adjacent cells in $\mathcal{C}$ that have at least one cell belonging to $\mathcal{C}^\star$. E_d, E_r and λ are respectively the unary data term, the pairwise potential and the weighting parameter between the two terms. A_k and $L_{kk'}$ are respectively the area of cell k, and the length of the common edge between cells k and k': they are introduced to normalize the energy with respect to the size of cells.

The intuition behind the data term is that (i) an empty cell is more likely to be ground with an elevation value of 0, and (ii) a conflict cell is more likely to be roof with an elevation value as close as possible to its inherited elevations:

$$E_d(x_k) = \begin{cases} 0 & \text{if } k \in \mathcal{C}_{empty} \text{ and } x_k = 0 \\ \min\{|x_k - z|_{z \in Z_k}\} & \text{else if } k \in \mathcal{C}_{conflict} \text{ and } x_k \neq 0 \\ \gamma & \text{otherwise} \end{cases} \tag{4.2}$$

where γ is a penalty for not respecting this intuition.

The pairwise potential is a generalized Potts model that decreases the penalty between two cells when their common edges projected in 3D back-project well into the images. As the pairs of cells are considered with different elevations x_k and $x_{k'}$, each pair has exactly two common edges in 3D: one at elevation x_k, the other at elevation $x_{k'}$. The pairwise term is expressed by

$$E_r(x_k, x_{k'}) = min(G^l(x_k) + G^l(x_{k'}), G^r(x_k) + G^r(x_{k'})) \cdot \mathbb{1}_{\{x_k \neq x'_k\}} \tag{4.3}$$

where $G^l(x_k)$ (respectively $G^r(x_k)$) is a back-projection measure of the common edge at elevation x_k into the left (respectively the right) image.

In practice, the back-projection measure is defined as the absolute value of one minus the scalar product between the image gradients and the gradients of the back-projected edge. To average the impacts of back-projection measures $G(x_k)$ and $G(x_{k'})$, a sum operation is defined to measure the back-projection quality of the common edges into one image. In theory, projections in two different perspectives are necessary for verifying the fidelity of 3D geometry. Ideally, both the two back-projections at x_k and $x_{k'}$ fit well the discontinuities in images if the elevation values

are accurate. However, in order to cover the case of incomplete contours of occluded areas, a *min* operation is applied to define the total back-projection measure of the left and the right images, instead of a median function. An explanation is shown in 4.5.

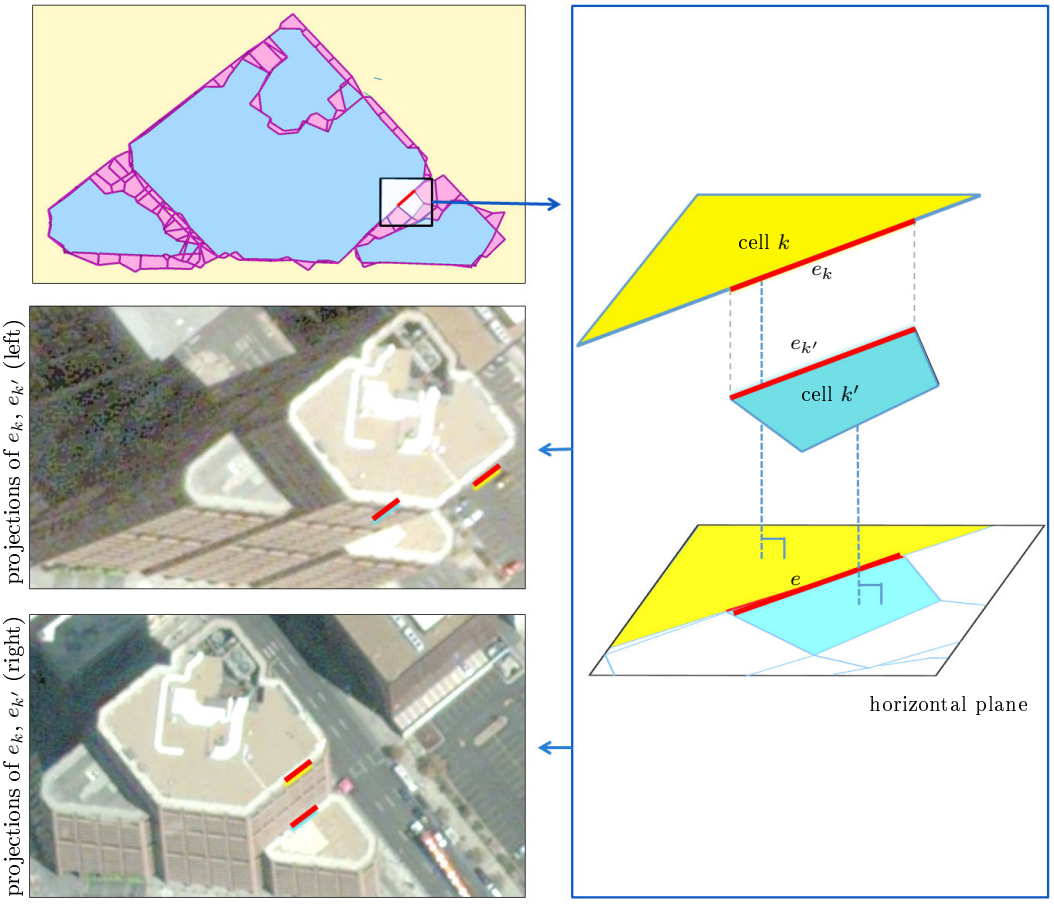


Figure 4.5: Back-projection measure E_r. The top left shows the projection of all enriched polygons on a horizontal plane; the right is a crop of cell k and cell k' projected in 3D space with label x_k and $x_{k'}$, and the line segment in red indicates the common edge e; the 2^{nd} and 3^{rd} row (left) describes the back-projections of the common edge e respectively with elevation x_k and $x_{k'}$ into the left and the right images.

Optimization. For efficiency reasons, the energy minimization is spatially decomposed into independent subproblems. The connected conflict and coherent cells are regrouped into clusters while allowing empty cells to be inside. Each cluster intuitively corresponds to a building or a building block. The α-β swap algorithm [BK04] is then operated over the set of conflict and empty cells of each cluster. Note that, for each cluster, the label set Z is restricted to the inherited elevations of its cells. Optionally, the optimization can be performed in parallel on each cluster.

Compact city model. The ground is represented in 3D by a mesh surface triangulated from the altitude estimates [BPD02]. From the optimal label configuration, roofs are inserted by simply elevated cells to their elevation label from the ground. The facade components are finally added by creating vertical facets between the adjacent cells with different labels.

4.4 Conclusion

This chapter presents a model fusion algorithm to extract optimal object contours from the two sets of enriched polygons produced by the joint classification in Chapter 3. The fusion method produces geometrically accurate 3D representations with high compactness by minimizing an energy function constrained with geometric accuracy and radiometric discontinuities. It gains robustness to low resolution and occlusion problems of satellite imagery by retrieving geometry and semantics simultaneously, and brings efficiency by operating on polygons.

Chapter 5

Experiments

The works in Chapter 2, Chapter 3 and Chapter 4 respectively describe the three main steps of the proposed automatic pipeline for city reconstruction from satellite images: polygonal partitioning, joint classification, and 3D model fusion. The polygonal partitioning algorithm decomposes images into convex polygons while preserving geometric shapes. The joint classification method relabels the polygonal partitions with semantic classes and elevation estimates simultaneously. Our model fusion algorithm generates geometrically accurate 3D models by optimizing object contours from two sets of enriched polygons. This automatic pipeline produces compact and geometric accurate 3D models for cities worldwide in a few minutes, with high scalability.

In this chapter, evaluations are presented to measure the efficiency and the performance of the automatic pipeline for 3D city reconstruction of different types of cities around the world. Input data acquisitions are stereo pairs from QuickBird 2, WorldView 2 and Pleiades satellite images with spatial resolution at 0.6, 0.5 and 0.5 meter respectively. All experiments are performed on an Intel Core i7 clocked at 2GHz.

5.1 Implementation details

The reconstruction pipeline is implemented in C++ using the Computational Geometry Algorithms Library (CGAL) [Cgal15] for manipulating geometric data structures in 2D and 3D, and the Geospatial Data Abstraction Library (GDAL) [GDA] for processing basic operations with satellite images. The Voronoi diagram structure in Section 2.5 is produced by the dual of the Delaunay triangulation.

For all satellite images in our experiments with a spatial resolution around $0.5m$, the parameter ε is fixed to 5. The cell decomposition in Section 4.3 is computed using a constrained Delaunay triangulation whose constrained edges correspond to the orthographic projection into the horizontal plane of the polygon edges of both partitions. The number of parameters is large, *i.e.,* 6, but this is the price to pay for a full pipeline combining semantic and geometric considerations in an unsupervised manner. In all the experiments, the weights of the two energies are fixed to $\beta_1 = 0.2$, $\beta_2 = 10$ and $\lambda = 2.5$, and the penalties to $\alpha = 0.05$ and $\gamma = 2$. The number of possible roof elevations is set to $n = 50$, except for US cities where skyscrapers requires increasing its value to 100.

The compactness of the output 3D model is mainly controlled by the weight of model fusion energy λ in Equation 4.3. It balances the compactness of models, the quality of contours, and the completeness (total area) of rooftops. A bigger value of λ leads the optimization solution to favor compactness of shapes and quality of contours, while losing patches that penalize regularity of shapes and back-projection quality. Figure 5.1 shows the impact of parameter λ on the reconstruction of a building.

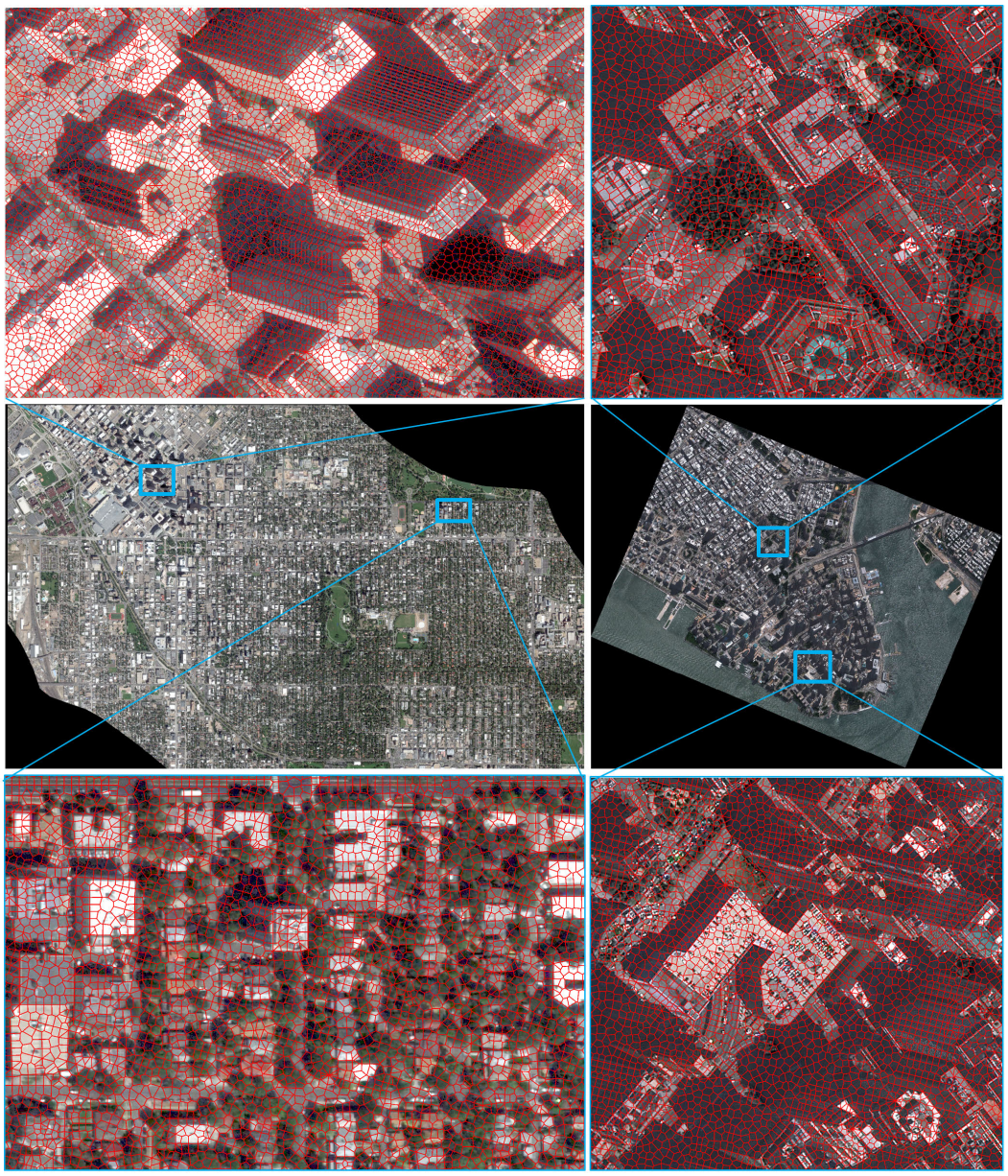
5.2 Qualitative evaluation

Experiments are conducted to evaluate the output qualities of the proposed algorithms from large-scale satellite images. Image polygonal partitioning and 3D reconstruction from large-scale cities are evaluated in terms of geometric shape preservation quality and altimetric accuracy.

Large-scale satellite images partitioning. Figure 5.2 shows an use case in which the algorithm characteristics are particularly attractive. Because of the scale and the geometric signature of the urban satellite images, the image partition well preserves the shape of buildings, in particular the facade and rooftop edges, as well as building corners. This knowledge is used in the geometry-aware joint classification of urban scenes.

Figure 5.1: Impacts of parameter λ. As λ increases (γ is fixed to 2.0), the model fusion method produces more compact object contours with higher back-projection quality. Optimization using only the unary data term in Equation 4.3, generates rough object contours without constraints on the regularity of geometric shapes(1^{st} column). A more compact and accurate outline of the building is captured by increasing λ to 5 (2^{nd} column). When increasing λ to 10, small geometric details can be obtained (yellow circle, 3^{rd} image), because the augmentation of back-projecting quality compensates the small increase of contour length. However, only small patches with regular shapes remain when set λ to an extremely large value (the 4^{th} column).

Reconstruction of buildings. 3D models produced by the proposed pipeline are georeferenced. If the model is accurate, the back-projection into the input stereo images of the roof edges from the output 3D model matches well the discontinuities of images. In a LOD1 representation, high quality back-projections of rooftops from both the left and right images prove the accuracy of geographic location and the precision of building elevations. Figure 5.3 displays a set of typical reconstruction of buildings of different types, such as a single roof with more than one height layer, building blocks, and freeform architecture. A visual comparison between the back-projected contours and the edges of rooftops in Figure 5.3 gives an average matching distance of less than 3 pixels, which means $1.5m$ at $0.5m$ spatial resolution. Due to the planar assumption of a polygonal region, curved roofs are approximated by a



Denver Manhattan, NY

Figure 5.2: Partition of large-scale satellite images. Our algorithm decomposes a 104Mpixel (respectively 39Mpixel) image of Denver (respectively Manhattan) into 0.8M (respectively 0.2M) convex polygons in a few minutes. The image partition nicely preserves the facade edges and rooftop junctions in spite of low image contrast (see close-ups).

step-like geometry.

5.3 Quantitative evaluation

Quantitative evaluation is conducted to show the geometric accuracy of the output 3D models. Comparison with DSM methods and LiDAR solutions shows the quality of our output models on buildings. To evaluate the accuracy of the output 3D model in a city-scale, an altimetric accuracy measurement is computed on Denver city using LiDAR data as ground truth.

Comparison with DSMs and LiDAR solutions. While there is no automatic algorithm producing compact and semantically-aware large-scale city models from satellite images, the output models are compared to traditional Digital Surface Models. The DSMs are generated from stereo matching following by structure recovery algorithms such as Voronoi clustering from the polygonal partitioning method (Chapter 2) and mesh simplification in [SLA15]. As shown in Figure 5.4, the output model better preserves the building structure while being semantically-aware and compact. The geometric accuracy of the proposed method is also measured in Figure 5.5 and is compared with accuracy of an airborne LiDAR based algorithm. Although the output is less accurate, the gap is relatively low given the contrast of data accuracy between airborne LiDAR and satellite imagery.

Altimetric accuracy on city reconstruction. An altimetric accuracy of the 3D city reconstruction on Denver is measured in Figure 5.6. The high quality Airborne LiDAR representation is considered as ground truth. Hausdorff distances are computed from each LiDAR point to the output 3D model from the proposed pipeline, and altimetric errors are displayed by projecting the Hausdorff distance values in to a XY grid. The proposed method generates compact 3D models within a reasonable average error to the ground truth. Most of the deviation constitutes to particular buildings missed because of insufficient number of elevation estimates, rough ground approximation due to elevated roads, and presence of trees that are not handled by the proposed algorithm.

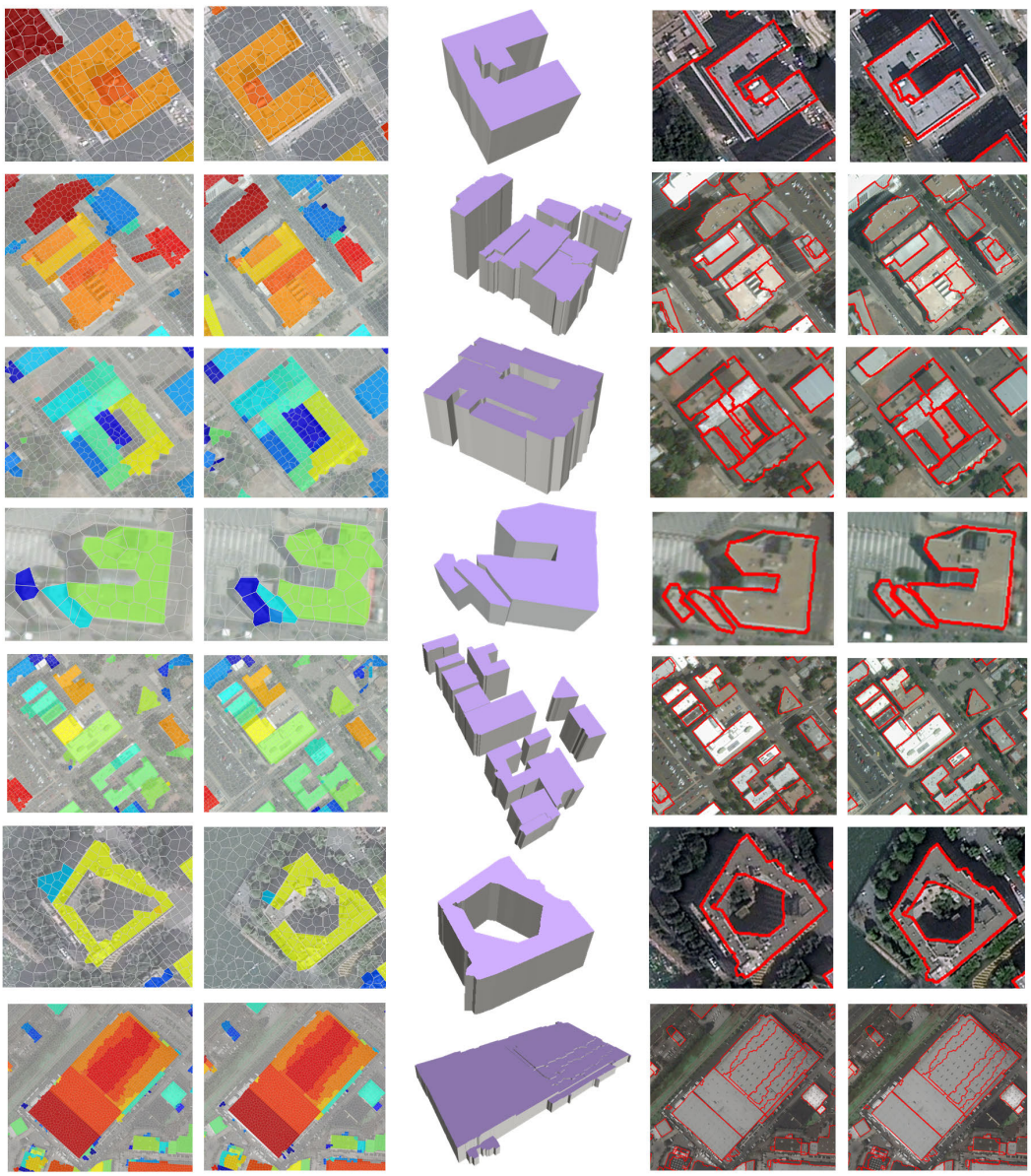


Figure 5.3: Reconstruction of buildings. On simple buildings (top rows), left and right enriched partitions (left columns) are relatively similar. For more complex buildings (middle rows), enriched partitions are more different: their fusion allows us to find a consensual 3D output model. With freeform architectural structures (bottom rows), curved roofs are roughly approximated by a step-like geometry. The back-projection into the input images of the roof edges from the output 3D model shows a good accuracy of both building elevations and contours (see red lines in right columns).

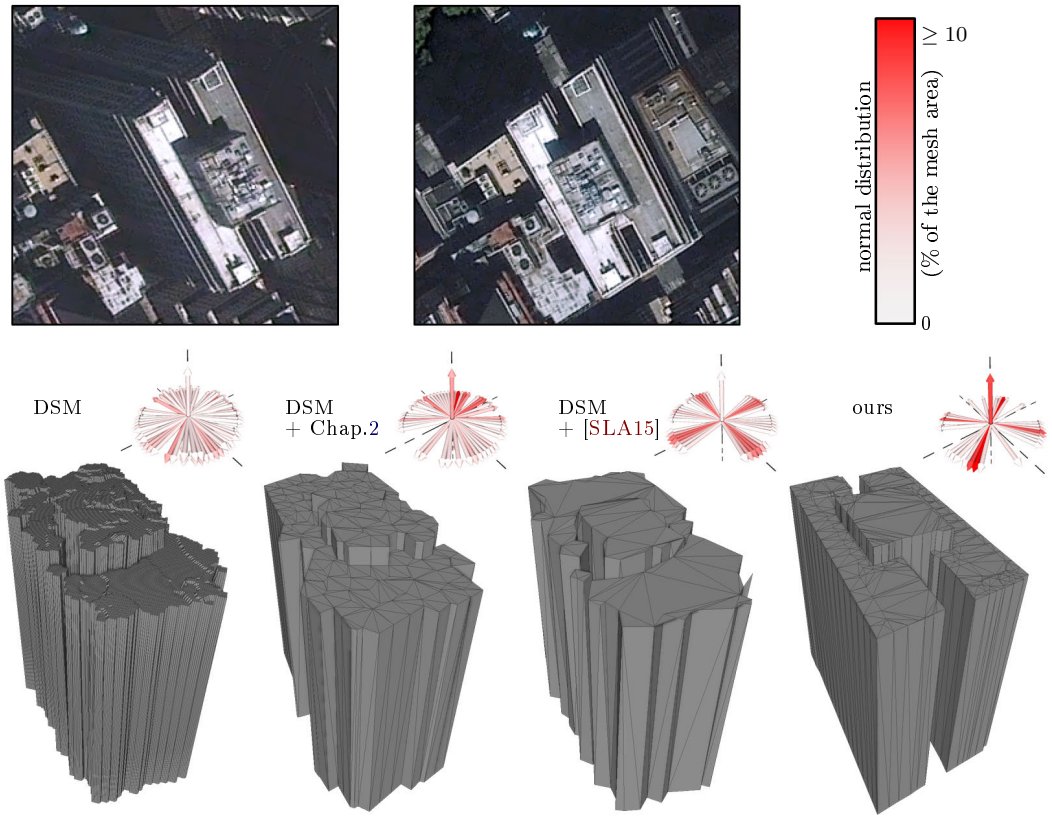


Figure 5.4: Comparison with Digital Surface Models. Traditional DSM derived from stereo matching [Hir08] at the pixel scale gives dense and structure-free 3D models. By postprocessing a DSM with Voronoi clustering proposed in Chapter 2 or with structure-aware mesh simplification [SLA15], more compact meshes are obtained, but the building structure cannot be not restored. The output model produced by the proposed pipeline is both compact and structure-aware (see the low number of principal directions in the distribution of output normals).

5.4 Robustness

The output models provide a faithful LOD1 representation of buildings, as illustrated in Figure 5.3. With the current satellite resolutions, a more detailed building representation such as LOD2 is not realistic. Cases that challenge our algorithm are small buildings, typically houses in residential areas, and the textureless and reflective objects which in general constitute an important challenge in stereovision.

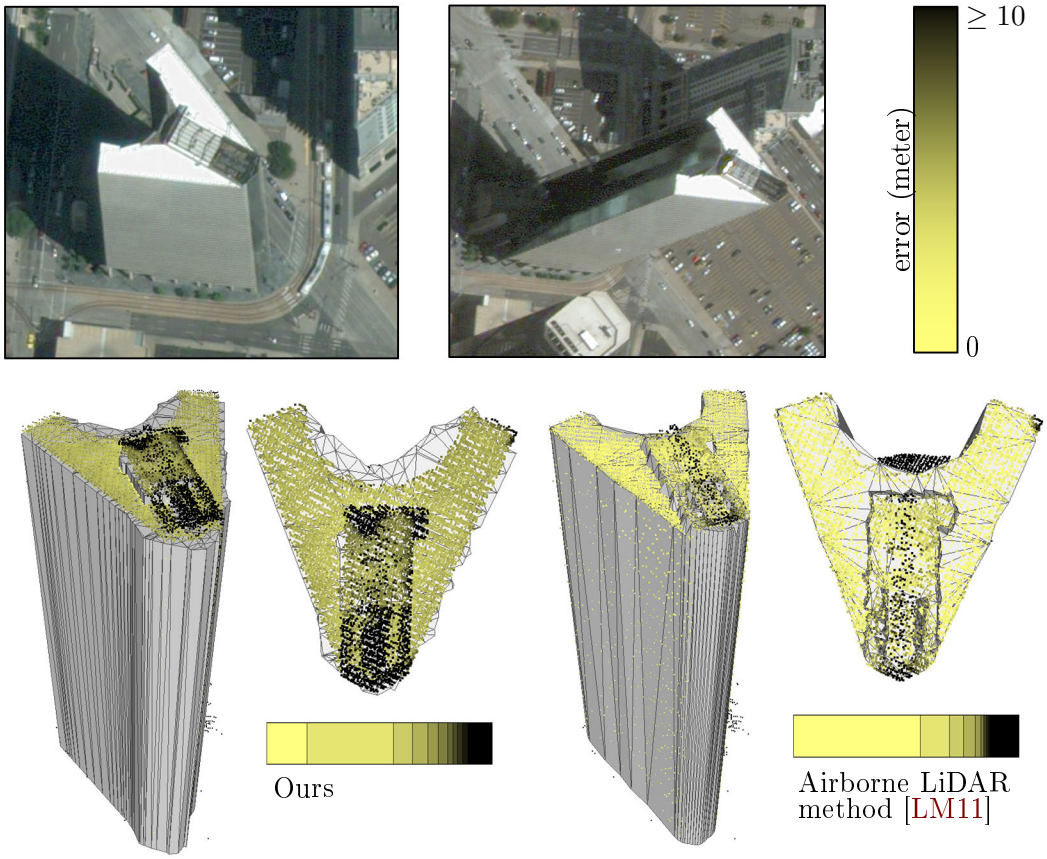


Figure 5.5: Geometric accuracy. Airborne LiDAR scans constitute precise measurements that can be used as ground truth to evaluate the geometric accuracy of the outputs of the proposed method (see the distribution of errors on the horizontal histogram). While a state-of-the-art airborne LiDAR method [LM11] produces more accurate results with a lower mean error to LiDAR points ($0.9m$ vs $1.7m$), the gap is relatively low given the difference of quality between the two types of inputs.

The proposed method can handle occlusions where some parts of buildings are invisible in one image, as shown in the top rows of Figure 5.7. However, severe roof occlusions where the occluded parts are large or a roof is only visible in one image, cannot be recovered, as shown in the bottom row of Figure 5.7.

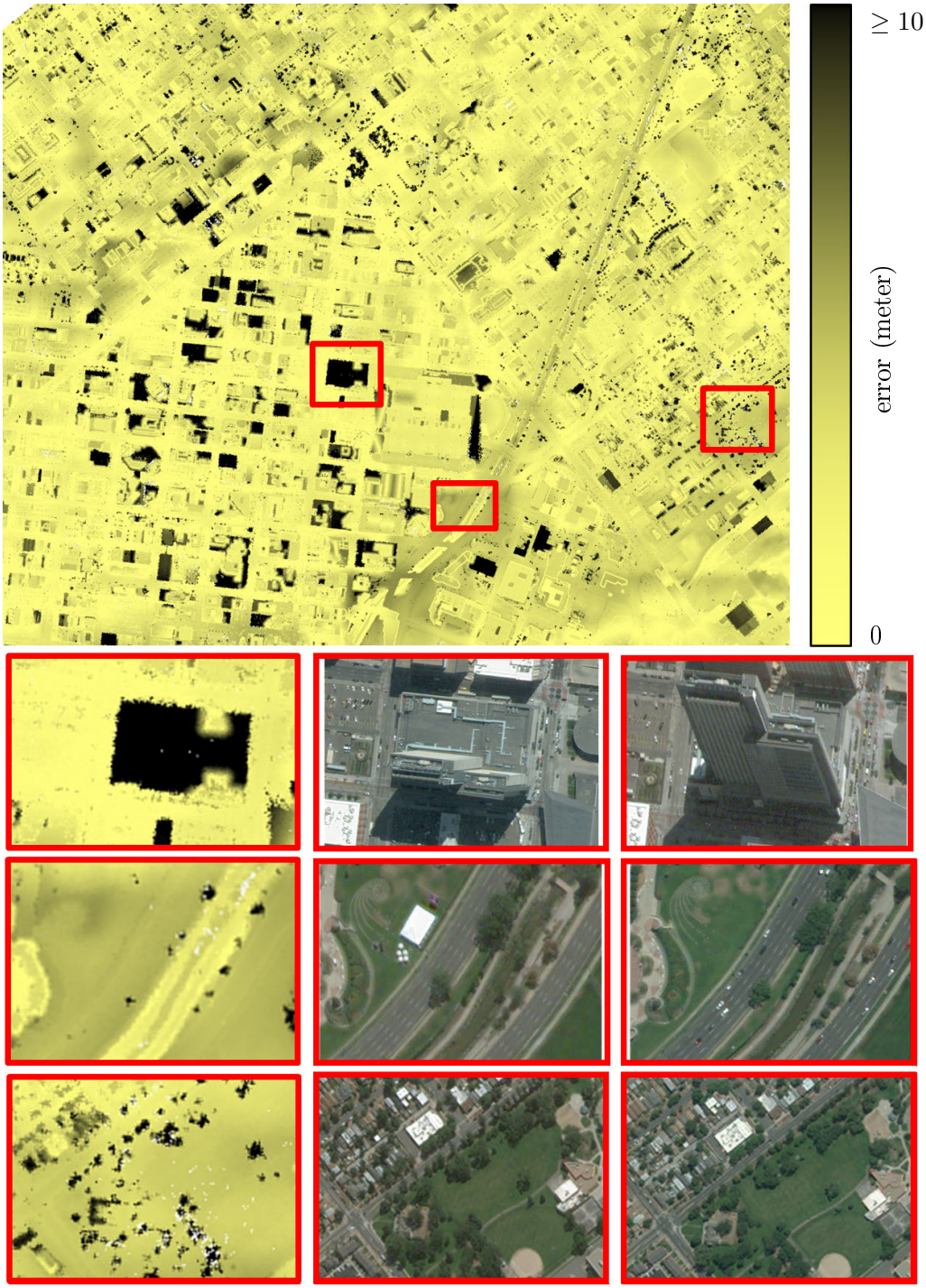


Figure 5.6: Altimetric accuracy on Denver. An altimetric error map (top) of our output 3D model on Denver (see Figure 1.1) is computed by, first, measuring the Hausdorff distance from each LiDAR point to the output 3D model, and, second, projecting the distance values into a XY grid. The closeups illustrate the main types of errors from insufficient number of elevation estimates, rough ground approximation due to elevated roads, and presence of trees.

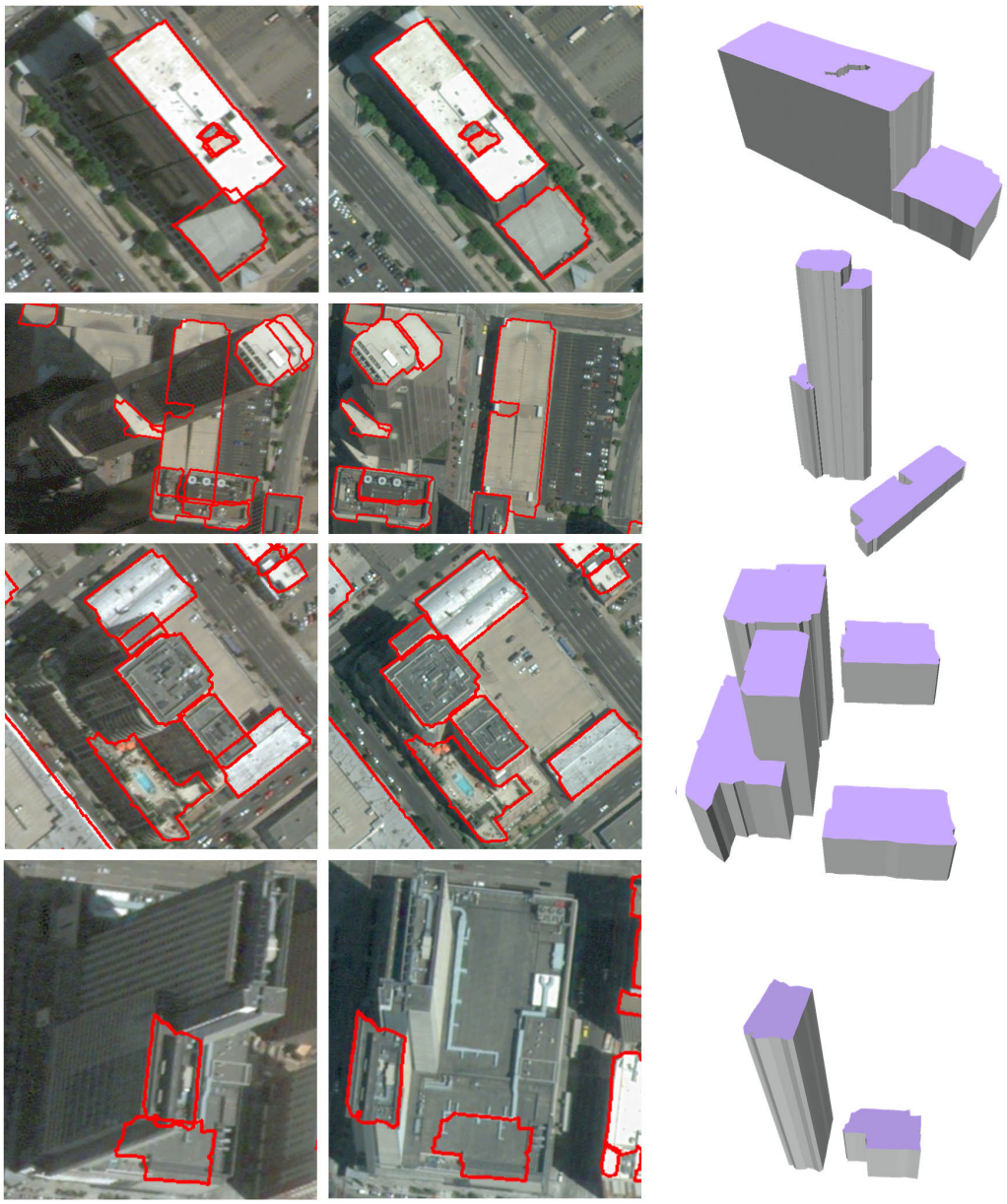


Figure 5.7: Robustness to occlusion areas. From the left to the right columns: the back-projections of the optimized contours on the left and right images, and the reconstructed 3D models. For roofs that are partially occluded in one image, the proposed pipeline can recover the occlusion areas (see the top three rows). However, large occlusion areas that lead to too few elevation estimates, may not be handled properly. In the bottom row, the highest part of the roof is missing because of the huge perspective differences induced by its non-flat structure, whereas its lowest (and largest) part is completely hidden in the left image, except one corner that is actually captured.

5.5 Scalability

The proposed pipeline has been tested on several cities presenting different urban landscapes, as shown in Figures 4.2 and 5.8. Dense downtowns in antique cities such as Alexandria, Egypt, are particularly challenging with narrow streets and small buildings massively connected. The proposed algorithm sometimes fails separating blocks in between these narrow streets as their width can be smaller than the size of our polygons. Business districts of US cities as Denver or New York are the opposite landscape: buildings are large, tall and fairly separated from each other. The proposed algorithm typically performs better on such areas. In terms of classification, buildings are globally well detected. One of the main reasons is because the proposed method does not rely on a pure radiometric description of buildings. At the scale of big cities, the radiometric variability of buildings is too high to draw likelihoods. Buildings can be missed when there are not enough elevation estimates. This situation is relatively marginal in practice: a visual comparison between our output 3D model of Denver and the building footprints of a cadastral map give us less than 5% of missed buildings and 14% of invalid buildings, i.e., buildings with at least 20% of their footprints missed or over-detected. Experiments of more large-scale city reconstruction are shown in Appendix 7.2, such as Melbourne harbor in Australia, San Francisco in the US, Madrid downtown in Spain, and Prague downtown in the Czech Republic.

5.6 Performance

Performance of the proposed automatic reconstruction pipeline are measured by several standard metrics including running times, memory peak, and output complexity. Impacts of several parameters are also evaluated in terms of performance.

Efficiency and complexity. To evaluate the performance of the proposed automatic pipeline, timings and complexity of output 3D models are given for different cities in Table 5.1. Input satellite images typically have around 30Mpixels. Each of the three steps of our method takes a few minutes from a typical stereo pair of

Figure 5.8: Reconstruction of cities. The proposed pipeline performs on different types of urban landscapes, including dense downtown (top left), antique city (bottom left), and US downtown (bottom right). Each model was obtained from one stereo pair of satellite images.

satellite images. For very dense cities, fusion is the most time-consuming step as the high density of buildings generates complex cell decompositions. For cities with more space in between the buildings such as New York City or Denver, fusion is quite fast. Running times for joint classification and elevation recovery, and polygonal partitioning do not depend on the urban landscape, but on the input image size. Overall, the use of compact and efficient geometric data structures allow us to have competitive timings with respect to airborne-based methods.

	New York City US	Denver US	Seoul South Korea	Alexandria Egypt
Polygonal partitioning	0.5 min	1.0 min	0.8 min	0.5 min
Joint classification	2.8 min	4.7 min	3.4 min	2.5 min
Fusion	1.5 min	2.8 min	13.7 min	29.2 min
Total time	4.8 min	8.5 min	17.9 min	32.2 min
Output complexity	0.23M	0.35M	0.89M	1.35M

Table 5.1: Running times and output complexity of the reconstruction pipeline. The output complexity refers to the number of triangular facets in the output 3D model. Note that the fusion step has been optimized sequentially on each building cluster.

Particularly, the polygonal partitioning algorithm performs well on big size images as shown in Table 5.2. Five minutes and 0.8Gb of memory are required for a 100Mpixel image. By contrast, the superpixel method ERS takes 39 minutes and 34Gb memory, and the released versions of SLIC and SEEDS are not able to handle such image size. Manipulating geometric shapes instead of pixels makes our algorithm particularly scalable. In terms of storage, our polygon partition can be saved in a very compact way as a planar graph where each node refers to a polygon.

Impact of parameters. Two key parameters, the partitioning parameter ε and the number of elevations n are evaluated to illustrate their impact on the reconstruction performance. As ε increases, images are partitioned into larger atomic regions and, thus, a lower number of convex polygons. But we are facing a trade-off, as shown in Figure 5.9. On one hand, a lower number of polygons brings computational efficiency; on the other hand, the higher sizes and the convexity of the polygons make the preservation of fine details in geometric shapes difficult. In the joint classification algorithm, n possible elevation values for the roofs are defined from the clustering of all elevation values in the scene. A greater value of parameter n means a larger label space of possible elevation values, which is more accurate in elevation estimation but is more time consuming for optimization. The curve of

	church (154Kpixels)	Manhattan (39.1Mpixels)	Denver (104Mpixels)
line extraction	36ms	29.9s	114.2s
consolidation	3ms	9.1s	107.4s
anchoring	3ms	2.7s	32.7s
homogenization	32ms	10.2s	48.4s
total time	72ms	51.9s	302.7s
memory peak	12.63Mb	372.20Mb	756.26Mb

Table 5.2: Performance of the polygonal partitioning algorithm on different image sizes (church from Figure 2.1, and Manhattan/Denver from Figure 5.2) in terms of running time and memory consumption. Note that the total execution times for Manhattan and Denver are different from those listed in Table 5.1 because here, the input images are the entire city scenes instead of the overlapped areas of a stereo pair.

parameter n is shown in Figure 5.10.

5.7 Limitations

The reconstruction pipeline has several limitations:

- The output 3D models of the automatic pipeline only contain three semantic labels (ground, roof and facade). The design of the algorithm is, however, flexible enough to account for new urban classes in future works.
- The limited quality of satellite images makes the reconstruction of small buildings, typically houses in residential areas, difficult.
- The proposed system is robust to occlusions of facades, ground and parts of roofs, but cannot handle severe roof occlusions where a roof is only visible in one image. The output LOD1 representation is also less accurate with freeform architectural roofs as domes or peaky structures. In such cases, roofs are approximated by a step-like geometry whose accuracy depends on the amount of elevation estimates.
- The polygonal partitioning is designed to partition images with a polygonal

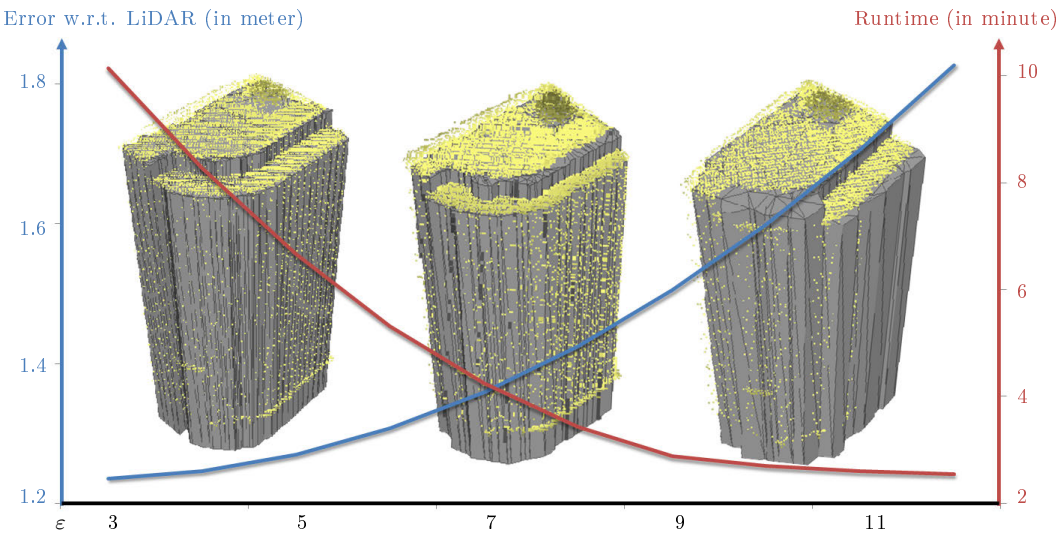


Figure 5.9: Impact of parameter ε. Using Airborne LiDAR scans as Ground Truth, geometric errors are measured by computing Hausdorff distances from each LiDAR point to the output 3D model. With a smaller ε, the output 3D models capture fine geometric shapes with higher accuracy, obtaining averagely lower geometric errors but longer runtime.

approximation of region boundaries that is usually relevant for man-made environments. The accuracy of the partitions is also dependent on the quality of the line-segments detected by the LSD [VGJMR10]. It produces accurate line-segments, however, still requires global regularization to get line-segment configuration of very high quality. In future works, the use of flexible geometric shapes that capture better freeform objects can be explored, and Quadrics or B-splines are potential solutions assuming Voronoi diagram can be built in non-Euclidean space that conforms to these shapes.

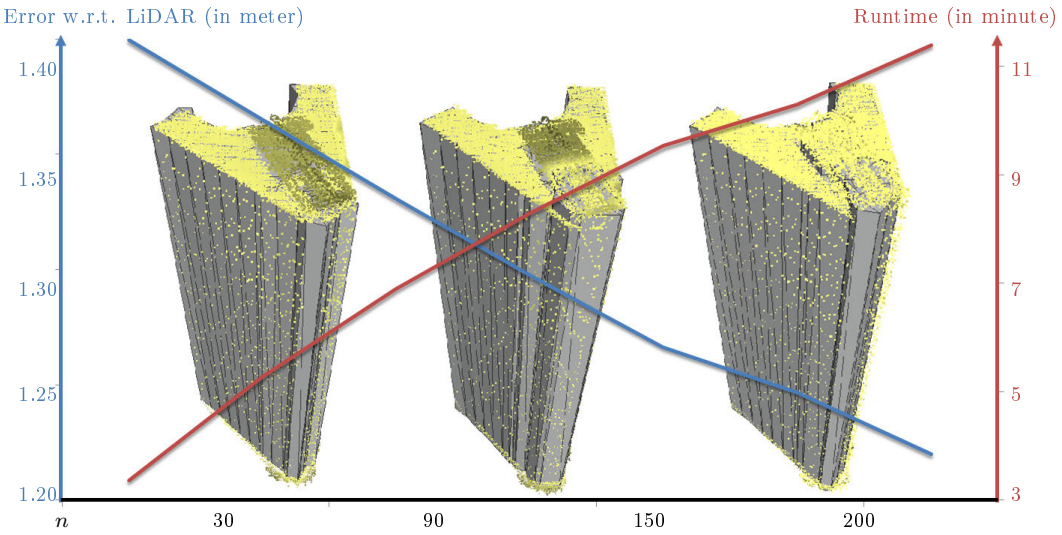


Figure 5.10: Impact of parameter n. The higher the value of n, the larger the label space $L = \{z_1, ..., z_n, other\}$ of the possible evaluation values. The elevation estimates of roofs contained in the 3D models obtained with a larger value of n get higher accuracy, at the cost of a loss in time efficiency.

Chapter 6

Conclusion

6.1 Summary

In this thesis, a fully automatic pipeline of 3D city reconstruction from satellite imagery has been proposed. Contributions are summarized, with a discussion of limitations.

Automatic city reconstruction. Due to the high diversity and complexity of urban contexts, a trade-off between reconstruction accuracy and algorithmic flexibility generally exists in many 3D city modeling methods. Approaches that rely on more predefined assumptions achieve higher accuracy of output city models, but lose flexibility to adapt to different types of urban scenes. For reconstruction from satellite imagery, the challenge is even greater because of its limited resolution, low image quality and long baseline problems. An automatic pipeline of urban city reconstruction from a satellite stereo pair is proposed in this thesis. It produces compact, semantically-aware and geometrically accurate 3D city models with high scalability and time efficiency by operating on polygonal atomic regions. Geometry and semantics are retrieved simultaneously in order to handle occlusion areas and low image quality robustly. Whereas the quality of the output models is not as accurate as airborne LiDAR solutions, the proposed method outclasses traditional DSM representations, and offers new perspectives in automatic city modeling.

Geometry-aware image partitioning. Image partitioning is a valuable technique to improve scalability and efficiency for large-scale city modeling from imagery data sources. Reducing radiometric redundancies by partitioning images into atomic regions brings advantages in terms of algorithmic complexity and spatial consistency.

In urban scenes, the most observed semantic objects are man-made structures with strong geometric signatures. A novel algorithm has been proposed in Chapter 2 to decompose images into convex polygons while preserving geometric shapes of urban objects. It provides geometric guarantees on connectivity and convexity of polygons and is suited well for man-made environments. With such a geometry-aware image partitioning technique, city modeling from imagery can be processed at the scale of polygons efficiently.

Joint exploration of semantics and elevations. Many city modeling strategies are either image-based or depth-based. The combination of them offers an interesting direction to explore. In this thesis, a joint classification algorithm is proposed to retrieve semantics and estimate elevations simultaneously through a global optimization. The 2D radiometric and 3D geometric clues are exploited in a common framework, gaining robustness to low image quality and occlusion problems. This joint strategy is not restricted to satellite imagery, but can be also applied to other stereo imagery applications.

Limitations. Several restrictions remind us that many challenges still remain in city reconstruction field.

Freeform-roof buildings. The proposed pipeline is not well adapted to the reconstruction of buildings with freeform roof shapes. First, the polygonal partitioning relies on detected line-segments and approximates curves by polylines. Regular geometric shapes are captured by adhering edges of polygons to the line-segments. This strategy is not adaptive to freeform shape preservation due to a loss of accuracy when represented with a limit number of straight edges. Secondly, the output LOD1 representation is less accurate with freeform architectural roofs such as domes or spires. In such cases, roofs are approximated by a step-like geometry whose accuracy depends on the number of elevation estimates.

Residential buildings. The limited quality of satellite images brings difficulties in the reconstruction of small buildings, typically houses in residential areas. It is a challenge for polygonal partitioning algorithm to preserve such small size shapes

of residential buildings by homogeneous polygons. Elevation data computed by the SGM from a satellite stereo pair is sparse and easy to miss small buildings. In addition, undesired objects such as vegetation and vehicles also make difficulties in accurate building detection.

Severe occlusion. Clues for recovering occluded areas mainly consist of radiometric information compensation from the two images and elevation values in their common parts. The proposed system, while robust to occlusions of facades, ground and pieces of roofs, cannot handle severe roof occlusions where a roof is only visible in one image.

6.2 Perspectives

Important challenges remain in 3D city modeling. The following areas are possible avenues for future works.

Modeling from higher resolution satellite imagery. With the fast advance of acquisition techniques, higher resolution satellite imagery becomes available. The newly launched WorldView 3 satellite provides images with a $0.31m$ spatial resolution, instead of $0.5m$ and $0.6m$ as used in this thesis. The gap between satellite imagery and aerial imagery or LiDAR scans still exists while becoming smaller and smaller in terms of spatial resolution. The long baseline problem may become the first challenge for adapting aerial-based approaches directly to city reconstruction from satellites. Big difference in perspectives brings difficulties in high quality dense matching that gives accurate details of urban objects. The second challenge may lie in the quality of satellite images. Higher resolution images provide richer information for reconstruction of geometric details of urban objects, but also introduce more noise and undesired objects.

Public domain and crowd-sourced data. During the last twenty years, different data sources such as airborne/satellite stereoscopic imagery or laser scanning are applied to tackle geometric modeling issues on urban scenes. The accessibility of

this data has become available to non-specialized people working in different communities. Even ubiquitous smartphones provide a large amount of photos that are available on the Internet, providing potential data source for 3D city reconstruction. This ever-expanding data, together with ever-increasing computational resources constitute a great opportunity to explore efficient solutions for 3D city reconstruction. First, public domain and crowd-sourced data provide a much wider coverage of worldwide cities compared the specialized data. Second, the constraints of single-source data are reduced such that existing modeling methods can be adapted to other data sources. This brings opportunities for multi-sourced, public, lower cost, and worldwide urban city reconstruction.

Learning characteristics of urban landscape. Many learning-based algorithms offer impressive performance on object recognition, data classification and semantic extraction. Prior knowledge learned from annotated or/and labeled samples can be integrated to enhance the robustness to undesired objects, e.g., vegetation and vehicles. Deep learning approaches do not require prior feature extraction but learn them directly from a training process. The learned networks is possible to describe the characteristics of urban landscapes well if the training dataset contains sufficient samples that depict various types of target semantic objects [Mni13]. Challenges still exist to adapt deep learning to urban reconstruction for the huge diversity and intra-class variation of semantic classes, and the difficulties in extensive annotation of training data. For high quality city reconstruction, one promising direction to explore is a hybrid system combining reliable learning-based semantic classification and 3D geometric shape reconstruction. Targeted objects retrieved from classification provides valuable clues in 2D perspectives for shape recovery in 3D. Another promising direction is learning-based procedural modeling, which learns to encode complex constraints using generative methods. The combination of learning approaches with grammar-splitting can be an interesting step to explore [TMT16].

Joint city modeling. The goal of semantic city modeling is to reconstruct an annotated 3D model, labeling objects as buildings, vegetation, ground, road networks and water areas, etc. Related literature shows a growing trend to fuse problems of segmentation, classification, and reconstruction altogether [FH+15]. Image segmentation plays an important role in reducing radiometric redundancy and computational complexity. Classification is useful for dealing with occlusions in modern urban reconstruction, particularly in long baseline stereo vision. It is still an open question how to best tackle segmentation, classification, and reconstruction from imagery in a common framework, aiming at automatic reconstruction of urban cities in an accurate and efficient way.

6.3 Conclusion (version française)

Cette dissertation a décrit un algorithme de reconstruction de villes en 3D à partir d'images satellites. Nous résumons ci-dessous nos contributions principales et en exposons les limitations.

Reconstruction automatique de villes. En raison de la diversité et de la complexité des grandes villes du monde, les algorithmes de modélisation urbaine en 3D sont généralement basés sur un compromis entre précision et flexibilité algorithmique. Les approches s'appuyant sur des hypothèses prédéfinies délivrent des résultats qui, pour certaines villes, offrent un bon niveau de précision, mais échouent à traiter convenablement certains types de scènes. Avec l'imagerie satellite, le défi est d'autant plus grand que les images sont de plus faible résolution et de moindre qualité, avec en outre, une plus longue droite épipolaire. Pour répondre à ces problèmes, cette dissertation a proposé un algorithme de reconstruction urbaine à partir de deux images stéréoscopiques. Les modèles 3D de villes en sortie de cet algorithme sont compacts, précis, et ses différents objets sont classés selon une sémantique particulière. Il se caractérise par un bon passage à l'échelle et une certaine rapidité, en opérant à une échelle géométrique élémentaire. La description géométrique des objets, et leur identité sémantique, sont extraites simultanément afin de gérer de

manière robuste les problèmes liés aux phénomènes d'occlusion et à la faible qualité des images. Bien que la qualité des modèles en sortie de l'algorithme ne soit pas aussi précise qu'avec les solutions basées sur des données LiDAR aériennes, la solution proposée surpasse les représentations DSM traditionnelles et ouvre de nouvelles perspectives dans la modélisation automatique de villes.

Partitionnement géométrique de l'image. Le partitionnement d'une image est une technique intéressante pour modéliser efficacement des villes à grande échelle et à partir de différentes sources de données. La réduction des redondances radiométriques par le partitionnement en régions atomiques procure de précieux avantages en termes de complexité algorithmique ou de cohérence spatiale. Au sein d'une scène urbaine, la plupart des objets observés sont des constructions humaines, avec des signatures géométriques fortes. Un algorithme de partitionnement d'une image en polygones convexes, préservant la forme géométrique des objets, a été décrit au chapitre 2. Celui-ci fournit d'importantes garanties géométriques sur la connectivité et la convexité des polygones, et s'adapte bien aux environnements créés par l'homme. Cette technique permet de traiter efficacement le problème de la reconstruction 3D en transformant l'image d'origine en simples polygones capturant la forme géométrique des objets.

Processus commun de classification sémantique et d'estimation des hauteurs. De nombreuses stratégies de modélisation sont soit basées sur une analyse des images, soit sur des cartes de profondeur. La combinaison des deux offre une piste intéressante de réflexion. Dans cette thèse, un algorithme est proposé pour classer sémantiquement les objets et estimer leur hauteur de façon simultanée, à l'aide d'un processus d'optimisation global. Les signaux radiométriques 2D et les formes géométriques 3D sont exploités dans un processus commun, conférant ainsi de la robustesse à notre algorithme face aux problèmes d'occlusion et de faible qualité des images. Cette double approche ne concerne pas seulement les problèmes liés à l'imagerie satellite, mais peut aussi être appliquée à d'autres problèmes en lien avec l'imagerie stéréoscopique.

Limites. Plusieurs difficultés rencontrées par cet algorithme nous rappellent que de nombreux défis doivent encore être relevés dans le domaine de la reconstruction urbaine.

Formes de toits libres. La chaîne de traitement proposée n'est pas adaptée à la reconstruction de bâtiments possédant des toits non plats. Premièrement, le partitionnement polygonal s'appuie sur les segments retournés par un algorithme de détection de contours, et approxime les formes incurvées par des courbes affines par morceaux. Les formes géométriques régulières sont capturées à l'aide de segments, positionnés le long des arêtes des polygones. Cette stratégie n'est pas adaptée aux formes libres, en raison d'une perte de précision, elle-même due à la représentation d'une courbe à l'aide d'un nombre fini de segments. Deuxièmement, le modèle LOD1 est peu précis avec des formes de toits non plates, telles que les dômes ou les spires. De tels objets sont en fait approximés à l'aide de courbes en escalier, dont la précision dépend du nombre d'estimations de hauteurs.

Quartiers résidentiels. La qualité limitée des images satellites pose problème pour la reconstruction de bâtiments de petite taille, typiquement les maisons des quartiers résidentiels. Il est ardu pour l'algorithme de partitionnement de préserver la forme de tels bâtiments dans des ensembles de polygones homogènes. La carte des disparités calculée par le SGM à partir de deux images est creuse, et il arrive souvent de manquer les bâtiments les plus petits. En outre, la présence d'objets non désirés tels que des arbres ou des véhicules rendent difficile la détection de bâtiments.

Occlusions. Les indices dont nous disposons pour retrouver les portions de bâtiments affectées par un phénomène d'occlusion consistent principalement en une compensation radiométrique et des estimations de hauteurs concernant les zones communes aux deux images d'un couple stéréo. Bien que robuste face aux occlusions concernant les façades, le sol ou certaines portions de toits, l'algorithme ne parvient pas à gérer les cas d'occlusions sévères, lorsqu'un toit de bâtiment n'est visible que sur une seule image.

Chapter 7

Appendix

7.1 Applications of polygonal partitioning

Some short experiments are conducted on two concrete vision problems, i.e., object polygonalization and corner detection, to illustrate the applicative potential of our algorithm.

Object polygonization. Starting from a polygonal partition computed by our algorithm, we used a binary Markov Random Field to label each polygon as inside or outside. The interface between inside and outside regions directly forms polygonal contours, contrary to existing object polygonization methods that require complex algorithms in practice, e.g., [SCF14]. The MRF formulation is based on a standard 2-term energy taking into account image consistency (radiometric information) and pairwise interactions (smoothness via a Potts model). Figure 7.1 illustrates the potential of this strategy compared to [SCF14] for extracting roof contours from aerial images.

Corner detection. We also tested our algorithm for detecting corner points in images, more precisely L- and Y-junctions. This can be done by a simple local analysis of the junction-anchors from a polygonal partition of our algorithm. In particular, we detect L-junctions (respectively Y-junctions) when a junction-anchor is generated from 2 line-segments (resp. 3 line-segments). As illustrated in Figure 7.2, the proposed method obtained a detection rate of similar order of magnitude than a specialized algorithm [XDG14] on both synthetic and real images.

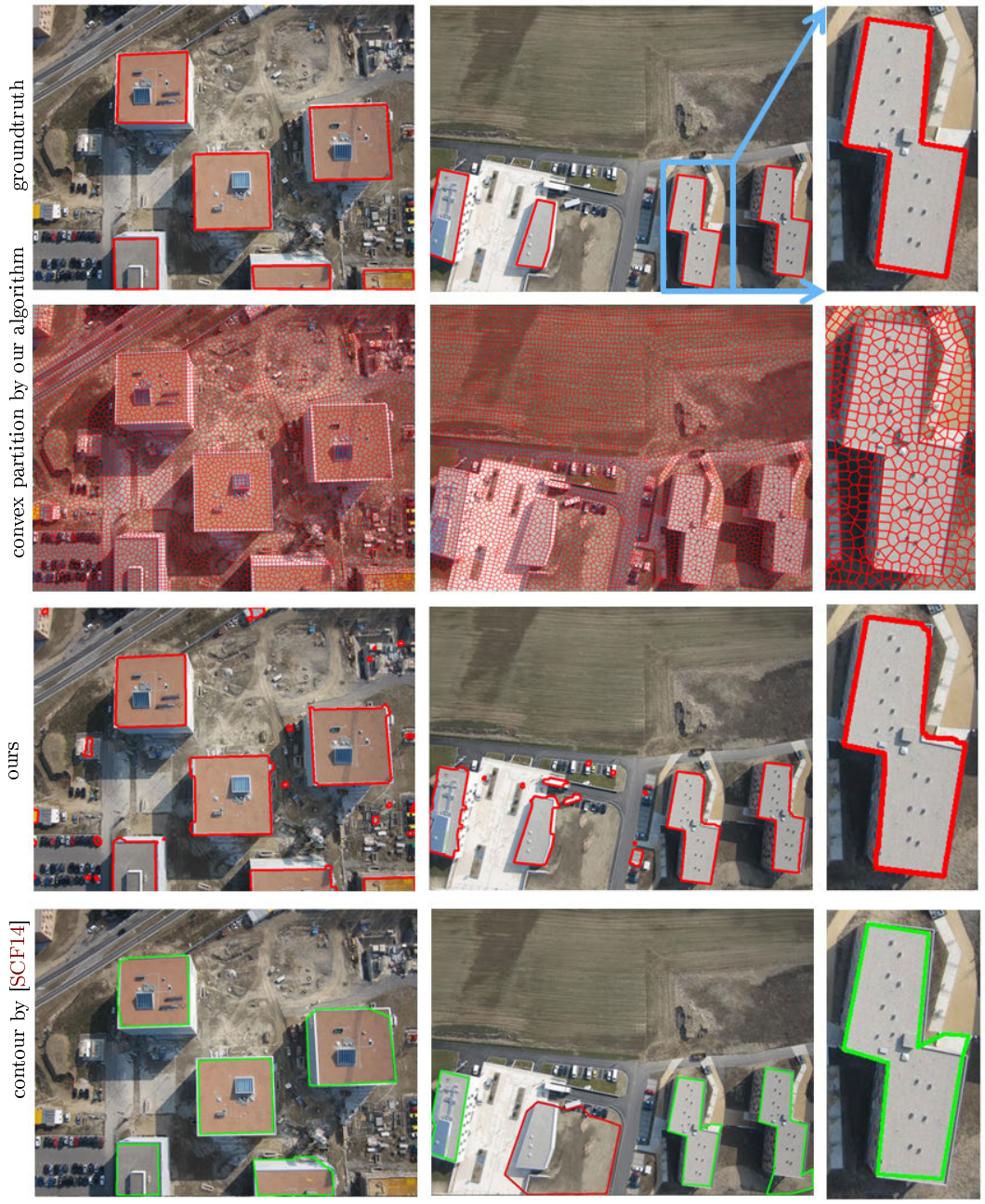


Figure 7.1: Object polygonization. Visual comparison of polygonal contour extraction by our method and by [SCF14] from aerial images. The class of interest corresponds to building roofs.

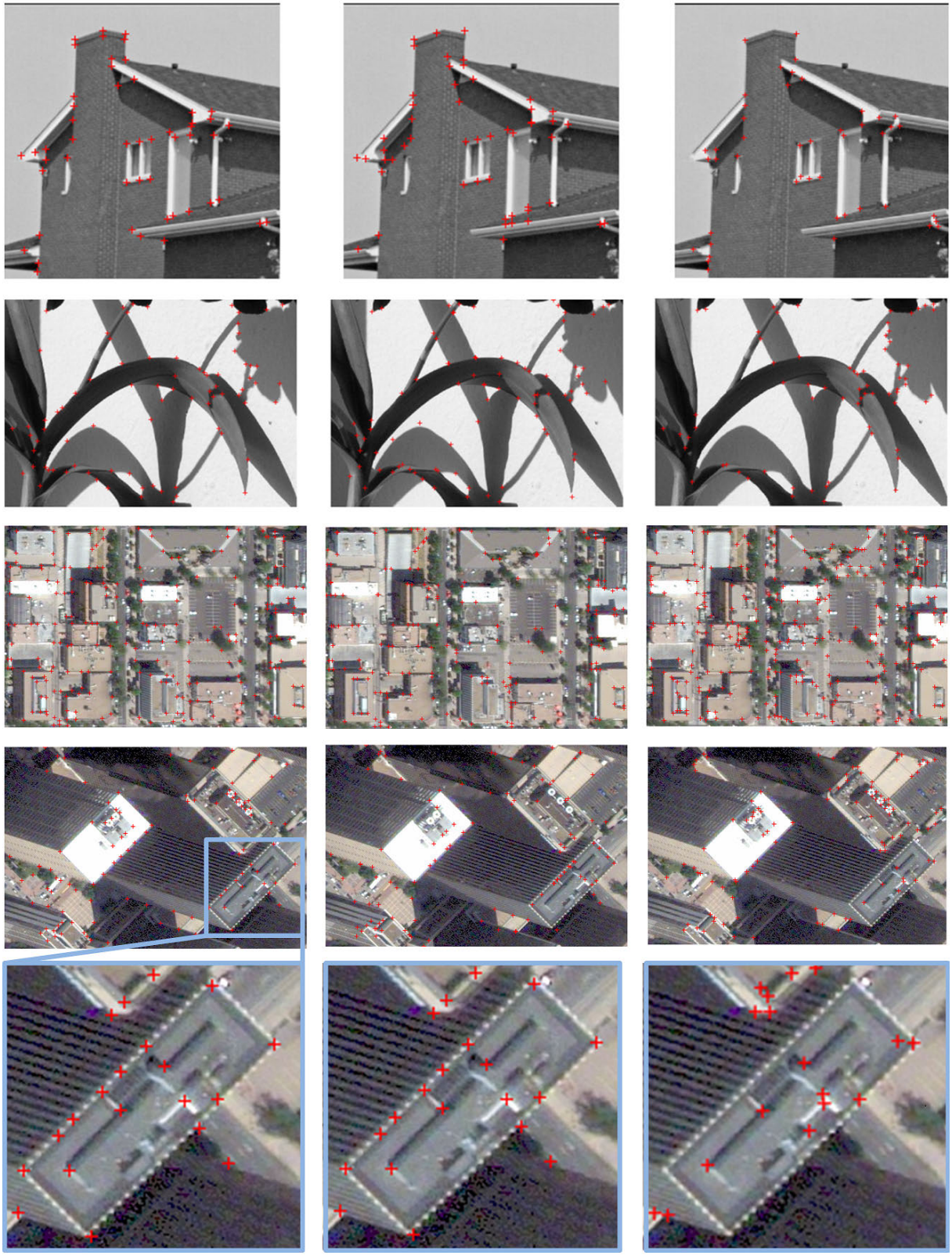


Figure 7.2: Corner detection. Visual comparison of corner detection by our method and ACJ [XDG14]. Our method produced similar results on both synthetic and real satellite images. In particular, for man-made objects as buildings, important corners of the roofs are correctly detected by our method.

7.2 Additional large-scale city reconstruction

Reconstruction of several additional cities is presented in this section. Different types of urban cities, such as Melbourne in Australia, San Francisco in the US, Madrid in Spain, and Prague in the Czech Republic, are reconstructed from satellite stereo pairs into compact and accurate 3D models, by the proposed automatic pipeline. The produced 3D models of these cities are respectively shown in Figures 7.3, 7.4, 7.5 and 7.6.

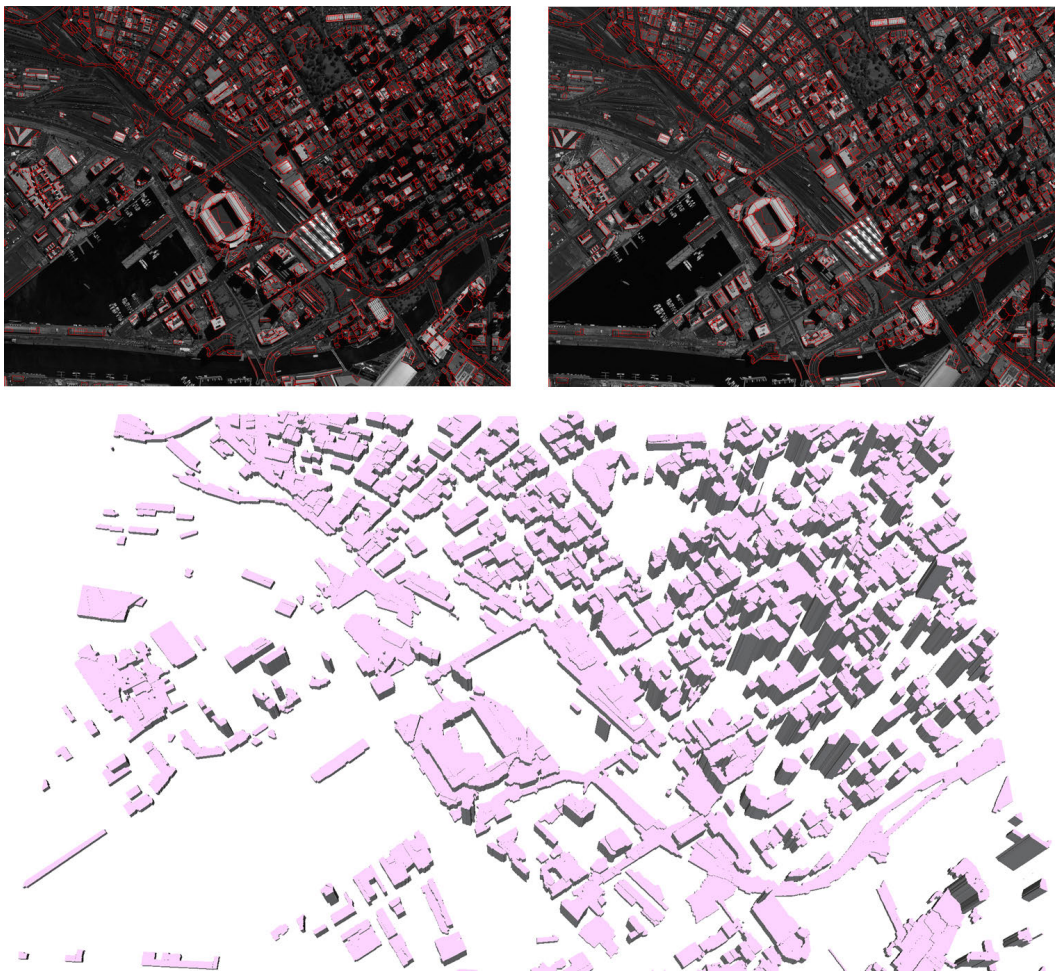


Figure 7.3: Reconstruction of Melbourne harbor, Australia. Our algorithm interprets elevated roads and bridges as buildings. Richer urban semantics are necessary to solve this problem. For evaluating the reconstruction accuracy with visual considerations, roof edges of the output 3D model have been back-projected into the input stereo satellite images (top, see red lines). This display scheme is also used for Figures 7.4 ,7.5 and 7.6.

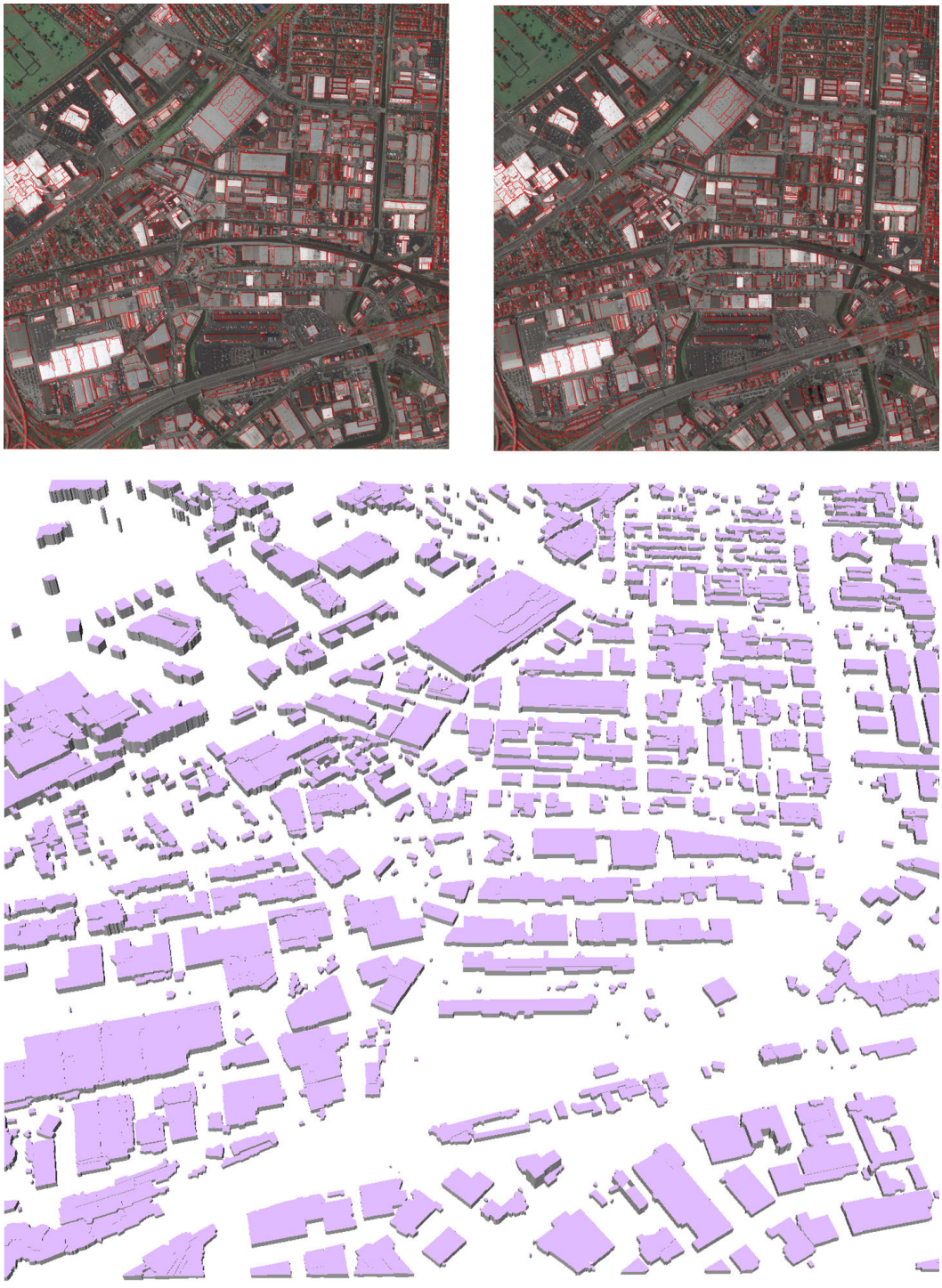


Figure 7.4: Reconstruction of an industrial district in San Francisco. Our LOD1 representation is well suited to industrial areas which are mainly composed of flat planar structures. Note that our algorithm is robust to building size variation, as we can see with both small houses and large industrial sites within the same scene.

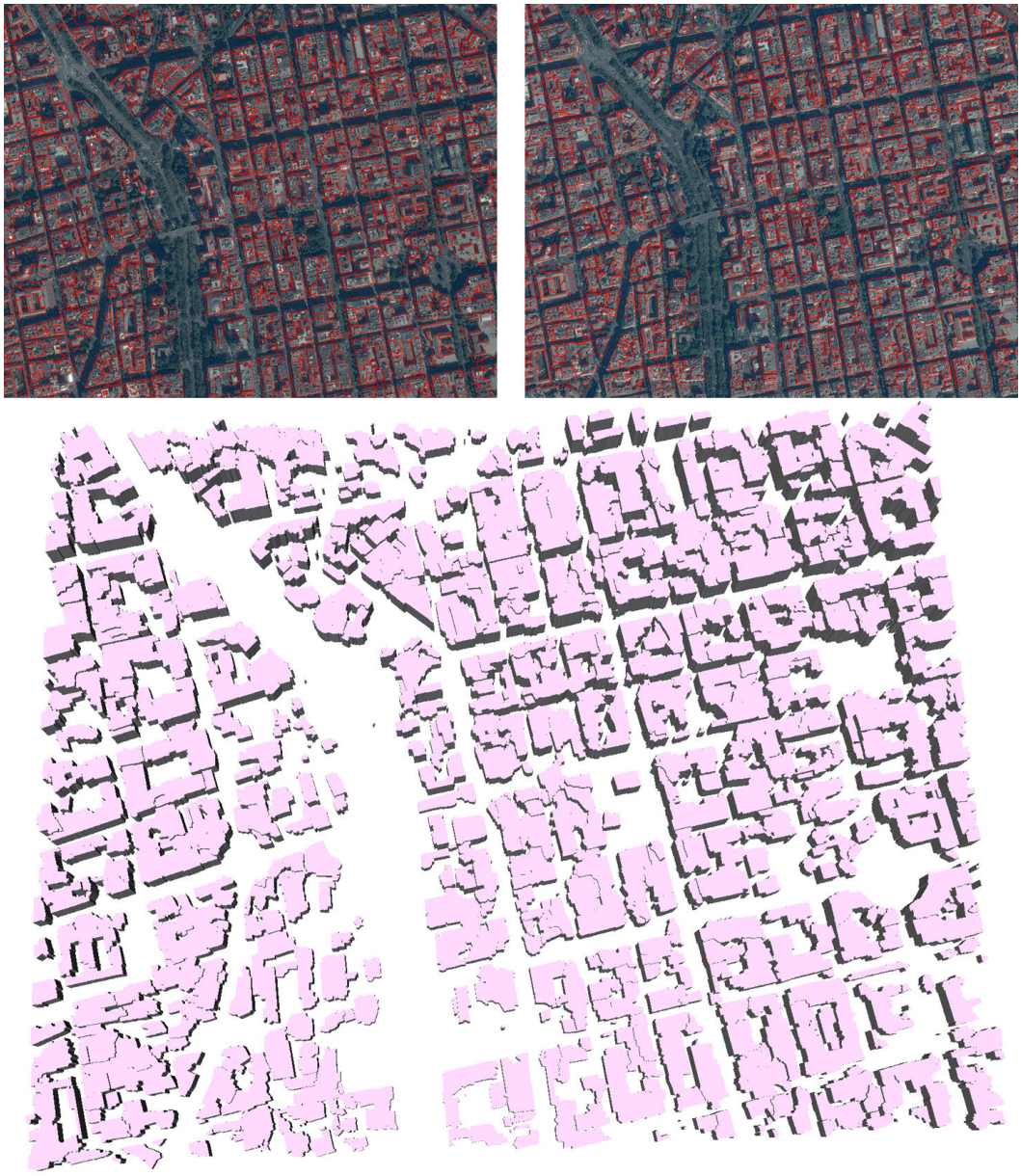


Figure 7.5: Reconstruction of Madrid downtown. Buildings of European cities as Madrid typically have sloping roofs that are better described by a LOD2 representation. Although our algorithm correctly extracts building blocks, high resolution airborne acquisition is recommended for capturing such roofs with better accuracy.

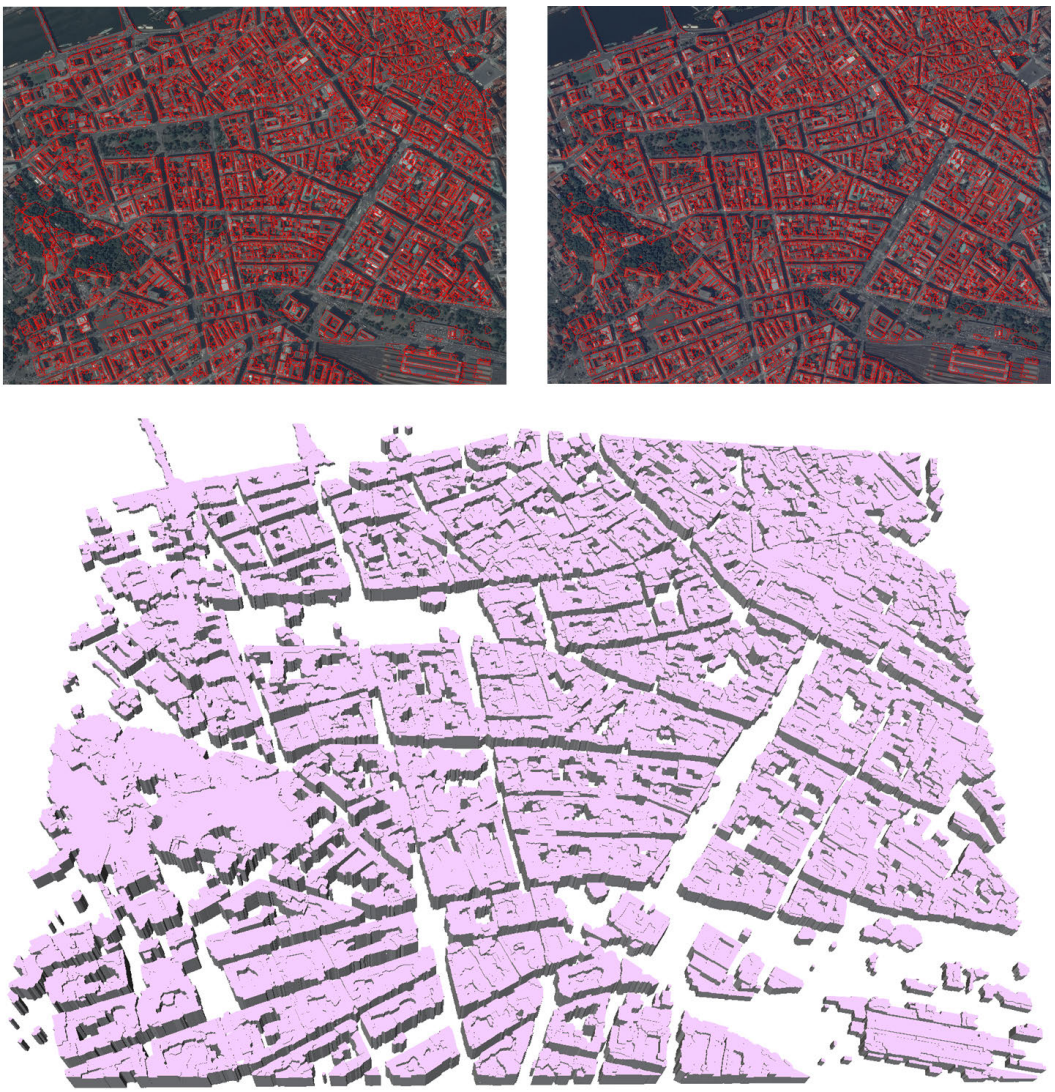


Figure 7.6: Reconstruction of Prague downtown. Although Prague exhibits urban characteristics similar to Madrid (Figure 7.5), its reconstruction is more challenging as (i) the road network (and by extension the building block layout) is more complex, and (ii) dense vegetation is more frequent and can easily be interpreted as buildings (see the middle left area on the output 3D model).

Bibliography

[AA10] H Gökhan Akçay and Selim Aksoy. Building detection using directional spatial constraints. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, pages 1932–1935. IEEE, 2010. (Cited on page 56.)

[AAC+06] Aseem Agarwala, Maneesh Agrawala, Michael Cohen, David Salesin, and Richard Szeliski. Photographing long scenes with multi-viewpoint panoramas. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 853–861. ACM, 2006. (Cited on page 6.)

[ADBW16] Daniel G. Aliaga, İlke Demir, Bedrich Benes, and Michael Wand. Inverse procedural modeling of 3d models for virtual worlds. In *ACM SIGGRAPH 2016 Courses*, SIGGRAPH '16, pages 16:1–16:316, New York, NY, USA, 2016. ACM. (Cited on page 23.)

[AFKH84] G. Asrar, M. Fuchs, E. Kanemasu, and J. Hatfield. Estimating absorbed photosynthetic radiation and leaf area index from spectral reflectance in wheat. *Agronomy Journal*, 76(2):300–306, 1984. (Cited on page 19.)

[AHL15] M. Ai, Q. Hu, and J. Li. A robust photogrammetric processing method of low-altitude uav images. *Remote Sensing*, 7(3):2302–2333, 2015. (Cited on page 14.)

[AM12] R. Attarzadeh and M. Momeni. Object-based building extraction from high resolution satellite imagery. In *ISPRS*, 2012. (Cited on page 18.)

[ARB07] Daniel G. Aliaga, Paul A. Rosen, and Daniel R. Bekins. Style grammars for interactive visualization of architecture.

IEEE Transactions on Visualization and Computer Graphics, 13(4):786–797, July 2007. (Cited on page 22.)

[ASS+12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11), 2012. (Cited on pages 39 and 49.)

[ASSS10] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *European conference on computer vision*, pages 29–42. Springer Berlin Heidelberg, 2010. (Cited on page 6.)

[BBW+09] M. Bokeloh, A. Berner, M. Wand, H.-P. Seidel, and A. Schilling. Symmetry detection using feature lines. *Computer Graphics Forum*, 2009. (Cited on page 21.)

[BFVG05] H. Bay, V. Ferrari, and L. Van Gool. Wide-baseline stereo matching with line segments. In *CVPR*, 2005. (Cited on page 57.)

[Bis06] C. Bishop. Pattern recognition and machine learning. In *Springer*, 2006. (Cited on page 16.)

[BK04] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9), 2004. (Cited on pages 63 and 75.)

[BPD02] C. Briese, N. Pfeifer, and P. Dorninger. Applications of the robust interpolation for dtm determination. In *Photogrammetric Computer Vision*, 2002. (Cited on pages 58 and 75.)

[BPJ07] O. Beeri, R. Phillips, and Hendrickson J. Estimating forage quantity and quality using aerial hyperspectral imagery for

northern mixed-grass prairie. *Remote Sensing of Environment*, 110(2):216–225, 2007. (Cited on page 19.)

[BRK+11] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo - joint stereo matching and object segmentation. In *CVPR*, 2011. (Cited on pages 16, 17, 27 and 57.)

[BSD09] M. Balzer, T. Schlomer, and O. Deussen. Capacity-constrained point distributions: a variant of lloyd's method. In *Proc. of Siggraph*, 2009. (Cited on pages 40 and 46.)

[BSRG14] András Bódis-Szomorú, Hayko Riemenschneider, and Luc Van Gool. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *CVPR*, 2014. (Cited on pages 26, 36 and 56.)

[BSRG15] András Bódis-Szomorú, Hayko Riemenschneider, and Luc Van Gool. Superpixel meshes for fast edge-preserving surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 7-12 June 2015. (Cited on pages 20, 26 and 27.)

[BVL93] A. Baddeley and M. Van Lieshout. Stochastic geometry models in high-level vision. *Journal of Applied Statistics*, 20(5-6), 1993. (Cited on page 40.)

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, November 2001. (Cited on page 13.)

[BWS10] Martin Bokeloh, Michael Wand, and Hans-Peter Seidel. A connection between partial symmetry and inverse procedural modeling. In *ACM SIGGRAPH 2010 Papers*, SIG-

GRAPH '10, pages 104:1–104:10, New York, NY, USA, 2010. ACM. (Cited on page 22.)

[CA09] G. Chen and Zakhor A. 2d tree detection in large urban landscapes using aerial lidar data. In *IEEE ICIP*, 2009. (Cited on page 19.)

[C. De Franchis] C. De Franchis. S2p. https://github.com/carlodef/s2p. (Cited on page 14.)

[CDSHD13] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *Trans. on Graphics*, 32(3), 2013. (Cited on page 36.)

[CF14] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 628–635, 2014. (Cited on page 70.)

[CFL13] Dengfeng Chai, Wolfgang Forstner, and Florent Lafarge. Recovering line-networks in images by junction-point processes. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, US, 2013. (Cited on pages 19 and 36.)

[Cgal15] Cgal. Computational Geometry Algorithms Library, 2015. http://www.cgal.org. (Cited on page 77.)

[CJSW01] H. Cheng, X. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects, pattern recognition. *Photogrammetric Engineering and Remote Sensing*, 34:259–2281, 2001. (Cited on page 16.)

[CL96] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *ACM SIGGRAPH*, 1996. (Cited on pages 24 and 25.)

[CLLS05] M. Charikar, E. Lehman, D. Liu, and A. Shelat. The smallest grammar problem. *IEEE Transactions on Information Theory*, 51(7):2554 – 2576, 2005. (Cited on page 21.)

[CLP10] A. Chauve, P. Labatut, and J. Pons. Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data. In *CVPR*, 2010. (Cited on pages 24 and 25.)

[Col96] Robert T. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, 1996. (Cited on page 14.)

[CPK+14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. (Cited on page 31.)

[CRZC04] Ana M Cingolani, Daniel Renison, Marcelo R Zak, and Marcelo R Cabido. Mapping vegetation in a heterogeneous mountain rangeland using landsat data: an alternative method to define and classify land-cover units. *Remote sensing of environment*, 92(1):84–97, 2004. (Cited on page 19.)

[CSAD04] David Cohen-Steiner, Pierre Alliez, and Mathieu Desbrun. Variational shape approximation. In *ACM SIGGRAPH*, 2004. (Cited on page 24.)

[CSF15] R. Cabezas, J. Straub, and J. Fisher. Semantically-aware aerial reconstruction from multi-modal data. In *ICCV*, 2015. (Cited on page 27.)

[CW11] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford Press, 2011. (Cited on pages 6 and 31.)

[DAB14] Ilke Demir, Daniel G. Aliaga, and Bedrich Benes. Proceduralization of buildings at city scale. In *Proceedings of the 2014 2Nd International Conference on 3D Vision - Volume 01*, 3DV '14, pages 456–463, 2014. (Cited on pages 21 and 22.)

[DAB15] Ilke Demir, Daniel G. Aliaga, and Bedrich Benes. Procedural editing of 3d building point clouds. In *ICCV*, 2015. (Cited on page 22.)

[DCNM14] M. Dang, D. Ceylan, B. Neubert, and Pauly M. Safe: Structure-aware facade editing. *Computer Graphics Forum*, 32(2):83–93, 2014. (Cited on page 21.)

[DF15] Carlo De Franchis. *Earth Observation and Stereo Vision*. Theses, Université Paris-Saclay, 2015. General Mathematics. (Cited on page 14.)

[DH06] D. Dunbar and G. Humphreys. A spatial data structure for fast poisson-disk sample generation. In *Proc. of Siggraph*, 2006. (Cited on page 40.)

[DHH11] M. Dubska, A. Herout, and J. Havel. Pclines – line detection using parallel coordinates. In *CVPR*, 2011. (Cited on page 39.)

[FBGS05] M. Fiocco, G. Bostrom, J. Gon¸calves, and V. Sequeira. Multisensor fusion for volumetric reconstruction of large outdoor areas. In *3DIM*, 2005. (Cited on pages 24 and 25.)

[FCSS09] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 1422–1429. IEEE, 2009. (Cited on page 26.)

[FFGG+10] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *European Conference on Computer Vision*, pages 368–381. Springer Berlin Heidelberg, 2010. (Cited on page 6.)

[FH+15] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. (Cited on page 97.)

[FP10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *PAMI*, 32(8):1362–1376, 2010. (Cited on page 14.)

[FSA05] C. Frueh, Jain S., and Zakhor A. Data processing algorithms for generating textured 3d building facade meshes from laser scans and camera images. *International Journal of Computer Vision*, 61(159), 2005. (Cited on page 27.)

[FVS12] B. Fulkerson, A Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Proc. of the European Conference on Computer Vision (ECCV)*, Firenze, Italy, 2012. (Cited on page 17.)

[FZ04] Christian Früh and Avideh Zakhor. An automated method for large-scale, ground-based city model acquisition. *International Journal of Computer Vision*, 60(1):5–24, 2004. (Cited on page 7.)

[GDA] GDAL. Geospatial Data Abstraction Library. http://www.gdal.org/. (Cited on page 77.)

[GDB85] K. Gallo, C. Daughtry, and M. Bauer. Spectral estimation of absorbed photosynthetically active radiation in corn canopies. *Remote Sensing of Environment*, 17(3):221–232, 1985. (Cited on page 19.)

[GDDA13] I. Garcia-Dorado, I. Demir, and D. Aliaga. Automatic urban modeling using volumetric reconstruction with surface graph cuts. *ComputersGraphics*, 37(7):896–910, 2013. (Cited on pages 24 and 25.)

[GFMP08] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *CVPR*, 2008. (Cited on page 28.)

[GH97a] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. In *SIGGRAPH*, 1997. (Cited on page 24.)

[GH97b] Rajiv Gupta and Richard I Hartley. Linear pushbroom cameras. *IEEE transactions on pattern analysis and machine intelligence*, 19(9):963–975, 1997. (Cited on page 12.)

[GH15] L. Gueguen and R. Hamid. Large-scale damage detection using satellite imagery. In *CVPR*, 2015. (Cited on pages 11 and 28.)

[GKF09] A. Golovinskiy, V. G. Kim, and T. A. Funkhouser. Shape-based recognition of 3d point clouds in urban environments. In *ICCV*, 2009. (Cited on page 15.)

[GP12] G. Groger and L. Plumer. Citygml – interoperable semantic 3d city models. *Journal of Photogrammetry and Remote Sensing*, 71, 2012. (Cited on pages 2 and 32.)

[GRB08] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. (Cited on page 17.)

[Gre95] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995. (Cited on page 26.)

[GW08] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, 3 edition, 2008. (Cited on page 29.)

[HCW+16] H. Hu, C. Chen, B. Wu, X. Yang, Q. Zhu, and Y. Ding. Texture-aware dense image matching using ternary census transform. In *ISPRS*, 2016. (Cited on page 14.)

[HDT+07] A. Van Den Hengel, A. Dick, T. Thormählen, B. Ward, and P. H. S. Torr. A shape hierarchy for 3d modelling from video. 63(9), 2007. (Cited on page 6.)

[Hig05] C. Higuera. A bibliographical study of grammatical inference. *Pattern Recognition*, 38(9):1332–1348, 2005. (Cited on page 21.)

[Hir08] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30(2), 2008. (Cited on pages 11, 14, 56, 57, 58 and 83.)

[HK92] P. Heermann and N. Khazenie. Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 30(1):81–88, 1992. (Cited on page 19.)

[HKDB13] Christof Hoppe, Manfred Klopschitz, Michael Donoser, and Horst Bischof. Incremental surface extraction from

sparse structure-from-motion point clouds. In *BMVC*, 2013. (Cited on page 25.)

[HKR+12] C. Hoppe, M. Klopschitz, M. Rumpler, A. Wendel, S. Kluckner, H. Bischof, and G. Reitmayr. Online feedback for structure-from-motion image acquisition. In *BMVC*, 2012. (Cited on page 25.)

[HS97] R. Hartley and T. Saxena. The cubic rational polynomial camera model. In *Image Understanding Workshop*, volume 649, 1997. (Cited on pages 10, 11 and 58.)

[HSU06] J. Hu, You S., and Neumann U. Integrating lidar, aerial image and ground images for complete urban building modeling. In *3DPVT*, 2006. (Cited on page 27.)

[HTC04] Yong Hu, Vincent Tao, and Arie Croitoru. Understanding the rational function model: methods and applications. *International Archives of Photogrammetry and Remote Sensing*, 20(6), 2004. (Cited on page 11.)

[HTP05] X. Hu, C. Tao, and B. Prenzel. Automatic segmentation of high-resolution satellite imagery by integrating texture, intensity and color features. *Photogrammetric Engineering and Remote Sensing*, 71:1399–1406, 2005. (Cited on pages 16 and 56.)

[HWA+10] Simon Haegler, Peter Wonka, Stefan Müller Arisona, Luc Van Gool, and Pascal Müller. Grammar-based encoding of facades. In *Proceedings of the 21st Eurographics Conference on Rendering*, EGSR'10, pages 1479–1487, Aire-la-Ville, Switzerland, Switzerland, 2010. Eurographics Association. (Cited on page 20.)

[HZ04] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. In *Cambridge University Press*, 2004. (Cited on pages 11 and 23.)

[HZC+13] Christian Haene, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, 2013. (Cited on pages 16, 17 and 27.)

[IMAW15] M. Ilcik, P. Musialski, T. Auzinger, and M. Wimmer. Layer-based procedural design of façades. *Computer Graphics Forum*, 34:205–216, 2015. (Cited on page 21.)

[Its15] Itseez. Open source computer vision library. `https://github.com/itseez/opencv`, 2015. (Cited on page 14.)

[IZB07] Arnold Irschara, Christopher Zach, and Horst Bischof. Towards wiki-based dense city modeling. In *2007 ieee 11th international conference on computer vision*, pages 1–8. IEEE, 2007. (Cited on page 6.)

[JA07] Secord J. and Zakhor A. Tree detection in urban regions using aerial lidar and image data. In *IEEE Geoscience and Remote Sensing Leters*, 2007. (Cited on page 19.)

[JLSW02] T. Ju, F. Losasso, S. Schaefer, and J. Warren. Dual contouring on hermite data. In *ACM SIGGRAPH*, 2002. (Cited on pages 24 and 25.)

[JP11] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, 2011. (Cited on pages 24 and 25.)

[JY12] Raquel Urtasun Jian Yao, Sanja Fidler. Describing the scene as a whole: Joint object detection, scene classifica-

tion and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. (Cited on page 17.)

[KBW+12] Javor Kalojanov, Martin Bokeloh, Michael Wand, Leonidas Guibas, Hans-Peter Seidel, and Philipp Slusallek. Microtiles: Extracting Building Blocks from Correspondences. *Computer Graphics Forum*, 31(5):1597–1606, 2012. (Cited on page 22.)

[KMRB09] Stefan Kluckner, Thomas Mauthner, Peter M Roth, and Horst Bischof. Semantic classification in aerial imagery by integrating appearance and height information. In *Asian Conference on Computer Vision*, pages 477–488. Springer Berlin Heidelberg, 2009. (Cited on page 56.)

[KOSPK16] P Kupidura, K Osińska-Skotak, and J Pluto-Kossakowska. Automatic approach to vhr satellite image classification. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 277–282, 2016. (Cited on pages 16 and 56.)

[KZ01] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, 2001. (Cited on page 13.)

[LA13] F. Lafarge and P. Alliez. Surface reconstruction through point set structuring. In *Proc. of Eurographics*, 2013. (Cited on page 43.)

[Laf14] Florent Lafarge. Some contributions to geometric modeling of urban environments. In *Habilitation thesis (HDR)*. University of Nice Sophia Antipolis, 2014. (Cited on page 3.)

[LC87] W. Lorensen and H. Cline. Marching cubes: A high res-

olution 3d surface construction algorithm. In *ACM SIGGRAPH*, 1987. (Cited on pages 24 and 25.)

[LC14] J. Li and M. Chen. On-road multiple obstacles detection in dynamical background. In *6th International Conference on Intelligent Human-machine Systems and Cybernetics*, 2014. (Cited on page 19.)

[LDH10] Matthew J Lato, Mark S Diederichs, and D Jean Hutchinson. Bias correction for view-limited lidar scanning of rock outcrops for structural characterization. *Rock mechanics and rock engineering*, 43(5):615–628, 2010. (Cited on page 8.)

[LDZPD08] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny. Building reconstruction from a single dem. In *CVPR*, 2008. (Cited on pages 25 and 26.)

[LGZ+13] Hui Lin, Jizhou Gao, Yu Zhou, Guiliang Lu, Mao Ye, Chenxi Zhang, Ligang Liu, and Ruigang Yang. Semantic decomposition and reconstruction of residential scenes from lidar data. *ACM Transactions on Graphics*, 32(4), 2013. (Cited on page 27.)

[LHK09] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009. (Cited on page 39.)

[LKB10] Florent Lafarge, Renaud Keriven, and Mathieu Bredif. Insertion of 3D-Primitives in Mesh-Based Representations: Towards Compact Models Preserving the Details. *IEEE Trans. on Image Processing*, 19(7):1683–1694, 2010. (Cited on pages 23 and 24.)

[LL05] S Lee and RG Lathrop. Sub-pixel estimation of urban land cover components with linear mixture model analysis and landsat thematic mapper imagery. *International Journal of Remote Sensing*, 26(22):4885–4905, 2005. (Cited on page 19.)

[LM11] Florent Lafarge and Clement Mallet. Building large urban environments from unstructured point data. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011. (Cited on pages 19, 27 and 84.)

[LMP01] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001. (Cited on page 27.)

[LPK07] P. Labatut, J. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *ICCV*, 2007. (Cited on page 25.)

[LRKT13] Lùubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Inference methods for CRFS with co-occurrence statistics. *International journal of computer vision*, 103(2):213–225, 2013. (Cited on page 17.)

[LSD10] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Optimal contour closure by superpixel grouping. In *ECCV*, 2010. (Cited on pages 35 and 70.)

[LSK+09] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi. Turbopixels: Fast superpixels using

geometric flows. *PAMI*, 31(12), 2009. (Cited on pages 39 and 49.)

[LSR+12] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *IJCV*, 100(2), 2012. (Cited on pages 15, 16, 27 and 57.)

[LT99] Peter Lindstrom and Greg Turk. Evaluation of memoryless simplification. *IEEE Trans. Vis. Comput. Graph.*, 5:98–115, 1999. (Cited on page 24.)

[LTRC11] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *CVPR*, 2011. (Cited on pages 39 and 49.)

[LW15] C. Li and M. Wand. Approximate translational building blocks for image decomposition and synthesis. *ACM Transactions on Graphics*, 34(5), 2015. (Cited on page 21.)

[LWC+11] Yangyan Li, Xiaokun Wu, Yiorgos Chrysanthou, Andrei Sharf, Daniel Cohen-Or, and Niloy J. Mitra. Globfit: Consistently fitting primitives by discovering global relations. *TOG*, 30(4), 2011. (Cited on pages 25 and 26.)

[LWM08] M. Lipp, P. Wonka, and Wimmer M. Interactive visual editing of grammars for procedural architecture. In *ACM SIGGRAPH*, 2008. (Cited on page 21.)

[LWWS15] C. Li, W. Wand, X. Wu, and H. Seidel. Approximate 3d partial symmetry detection using co-occurrence analysis. In *the 25th annual ACM symposium on User interface software and technology*, 2015. (Cited on pages 21 and 22.)

[LZMC12] Z. Li, Wu Z.-M., and S.-F. Chang. Segmentation using

superpixels: A bipartite graph partitioning approach. In *CVPR*, 2012. (Cited on page 35.)

[MCA+16] Blaha M., Vogel C., Richard A., Wegner J.D., Pock T., and Schindler K. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016. (Cited on page 17.)

[MCL+14] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. 2014. (Cited on page 17.)

[MFTM01] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. (Cited on page 49.)

[MGP06] N. Mitra, L. Guibas, and M. Pauly. Partial and approximate symmetry detection for 3d geometry. *ACM Transactions on Graphics*, 25(3):560–568, 2006. (Cited on page 22.)

[MK10] B. Micusik and J. Kosecka. Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision*, 89(1):106–119, 2010. (Cited on page 26.)

[MMWV11] Markus Mathias, Anđelo Martinović, Julien Weissenberg, and Luc Van Gool. Procedural 3d building reconstruction using shape grammars and detectors. In *3DIMPVT*, pages 304–311, 2011. (Cited on page 22.)

[MMWVG12] A. Martinovic, M. Mathias, J. Weissenberg, and

L. Van Gool. A three-layered approach to facade parsing. In *ECCV*, 2012. (Cited on page 21.)

[Mni13] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. (Cited on pages 56 and 96.)

[MPM+14] Oliver Mattausch, Daniele Panozzo, Claudio Mura, Olga Sorkine-Hornung, and Renato Pajarola. Object detection and classification from large-scale cluttered indoor scans. *Computer Graphics Forum*, 33(2):11–21, 2014. (Cited on page 16.)

[MSS+08] B. Matei, H. Sawhney, S. Samarasekera, J. Kim, and R. Kumar. Building segmentation for densely built urban regions using aerial lidar data. In *3CVPR*, 2008. (Cited on page 25.)

[MTCA16a] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Fully convolutional neural networks for remote sensing image classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016, Beijing, China, July 10-15, 2016*, pages 5071–5074, 2016. (Cited on page 31.)

[MTCA16b] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. High-resolution semantic labeling with convolutional neural networks. *CoRR*, abs/1611.01962, 2016. (Cited on page 31.)

[MTCA16c] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Recurrent neural networks to enhance satellite image classification maps. *CoRR*, abs/1608.03440, 2016. (Cited on page 31.)

[MVG13] A. Martinovic and L. Van Gool. Bayesian grammar learning for inverse procedural modeling. In *CVPR*, 2013. (Cited on page 21.)

[MVGV09] T. Moons, L. Van Gool, and M. Vergauwen. Reconstruction from multiple images part 1: Principles. *Foundations and Trends in Computer Graphics and Vision 4*, 4:287–404, 2009. (Cited on page 11.)

[MWA+13] Przemyslaw Musialski, Peter Wonka, Daniel G. Aliaga, Michael Wimmer, Luc van Gool, and Werner Purgathofer. Understanding the rational function model: methods and applications. *Computer Graphics Forum*, 32(6), 2013. (Cited on pages 4, 11, 13, 20, 26 and 31.)

[MWH+06] P. Muller, P. Wonka, S. Haegler, A. Ulmer, and L. Van Gool. Procedural modeling of buildings. In *SIGGRAPH*, 2006. (Cited on page 20.)

[MWL12] A. Ma, W. Wang, and S. Liu. Extracting roads based on retinex and improved canny operator with shape criteria in vague and unevenly. *Journal of Applied Remote Sensing*, 6(23):1–14, 2012. (Cited on page 19.)

[MZC05] S D Mayunga, Y Zhang, and J Coleman. Semi-automatic building extraction utilizing quickbird imagery. In *Proc. Int. Arch. Photogramm. Remote Sens.*, 2005. (Cited on page 18.)

[MZWLS15] JA Montoya-Zegarra, JD Wegner, L Ladickỳ, and K Schindler. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):127, 2015. (Cited on page 17.)

[MZWVG07] P. Mueller, G. Zeng, P. Wonka, and L. Van Gool. Image-based procedural modeling of facades. In *SIGGRAPH*, 2007. (Cited on page 20.)

[NMW97a] C. Nevill-Manning and I. Witten. Compression and explanation using hierarchical grammars. *Journal of Artificial Intelligence Research*, 40(2/3):103–116, 1997. (Cited on page 21.)

[NMW97b] C. Nevill-Manning and I. Witten. Identifying hierarchical strcture in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 7:67–82, 1997. (Cited on page 21.)

[OBSC00] A. Okabe, B. Boots, K. Sugihara, and S.N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley-Blackwell, 2000. (Cited on page 40.)

[Opentopography13] Opentopography. Opentopography webpage, 2013. http://www.opentopography.org. (Cited on page 8.)

[OTGB11] J. Oh, C. Toth, and D. Grejner-Brzezinska. Automatic georeferencing of aerial images using stereo high-resolution satellite images. *Photogrammetric Engineering & Remote Sensing*, 77(11):1157–1168, 2011. (Cited on page 11.)

[PC13] D. Poli and I. Caravaggi. 3d modeling of large urban areas with stereo VHR satellite imagery: lessons learned. *Natural Hazards*, 68(1), 2013. (Cited on pages 11 and 31.)

[PDP06] M. Pierrot-Deseilligny and N. Paparoditis. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from spot5-hrs stereo imagery. 36(1/W41), 2006. (Cited on page 14.)

[PMW+08] M. Pauly, N. Mitra, J. Wallner, H. Pottmann, and L. Guibas. Discovering structural regularity in 3d geometry. In *ACM SIGGRAPH*, 2008. (Cited on page 21.)

[PNF+08] M. Pollefeys, D. Nister, J. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. 78(23):143–167, 2008. (Cited on page 26.)

[PSP16] Andrea Cohen Pablo Speciale, Martin R. Oswald and Marc Pollefeys. A symmetry prior for convex variational 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016. (Cited on pages 20, 21 and 22.)

[PTN09] N. Plath, M. Toussaint, and S. Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. (Cited on pages 17 and 56.)

[PVGV+04] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004. (Cited on page 6.)

[PY09] C. Poullis and S. You. Automatic reconstruction of cities from remote sensor data. In *CVPR*, 2009. (Cited on page 25.)

[PZ14] V. Pawar and M. Zaveri. Graph based k-nearest neighbor minutiae clustering for fingerprint recognition. In *10th*

International Conference on Natural Computation, 2014. (Cited on page 19.)

[RBF12] X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: Features and algorithms. In *CVPR*, 2012. (Cited on page 36.)

[Ren07] X. Ren. Learning and matching line aspects for articulated objects. In *CVPR*, 2007. (Cited on page 39.)

[RHM02] H. Ruther, M. Hagai, and E. Mtalo. Application of snakes and dynamic programming optimization in modeling of buildings in informal settlement areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56:269–282, 2002. (Cited on page 18.)

[RJRO13] M. Reso, J. Jachalsy, B. Rosenhahn, and J. Ostermann. Temporally consistent superpixels. In *ICCV*, 2013. (Cited on page 39.)

[RvDHV06] T Rabbani, F van Den Heuvel, and G Vosselman. Segmentation of point clouds using smoothness constraint. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(5):248–253, 2006. (Cited on page 16.)

[RWFH12] M. Rothermel, K. Wenzel, D. Fritsch, and N. Haala. Sure: Photogrammetric surface reconstruction from imagery. In *Proceedings LC3D Workshop*, 2012. (Cited on page 14.)

[SAEAGS11] Ibrahim F Shaker, Amr Abd-Elrahman, Ahmed K Abdel-Gawad, and Mohamed A Sherief. Building extraction from high resolution space images in high density residential areas in the great cairo region. *Remote Sensing*, 3(4):781–791, 2011. (Cited on page 18.)

[SCD+06] S. Seitz, B. Curless, J. Diebel, D. Schartstein, and R. Szeliski. Stereo reconstruction algorithms. 2006. (Cited on page 23.)

[SCF14] X. Sun, M. Christoudias, and P. Fua. Free-shape polygonal object localization. In *ECCV*, 2014. (Cited on pages 36, 70, 101 and 102.)

[SFS12] A. Schick, M. Fischer, and R. Stiefelhagen. Measuring and evaluating the compactness of superpixels. In *ICPR*, 2012. (Cited on pages 39 and 49.)

[SGSS08] Noah Snavely, Rahul Garg, Steven M Seitz, and Richard Szeliski. Finding paths through the world's photos. *ACM Transactions on Graphics (TOG)*, 27(3):15, 2008. (Cited on page 6.)

[SHFH11] C. Shen, S. Huang, H. Fu, and S. Hu. Adaptive partitioning of urban facades. In *ACM SIGGRAPH Asia*, 2011. (Cited on page 21.)

[Sim11] C. Simler. An improved road and building detector on vhr images. In *International Geoscience and Remote Sensing Symposium*, 2011. (Cited on page 19.)

[SJC08] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. (Cited on pages 15 and 16.)

[SLA15] D. Salinas, F. Lafarge, and P. Alliez. Structure-Aware Mesh Decimation. *Computer Graphics Forum*, 34(6), 2015. (Cited on pages 24, 81 and 83.)

[SLS08] J. Shen, X. Lin, and Y. Shi. Knowledge-based road extraction from high resolution remotely sensed imagery. In *2008 Congress on Image and Signal Processing*, 2008. (Cited on page 19.)

[SS02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002. (Cited on pages 23 and 56.)

[SSS06] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, pages 835–846. ACM, 2006. (Cited on page 6.)

[SSS+08] Sudipta N Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. Interactive 3d architectural modeling from unordered photo collections. In *ACM Transactions on Graphics (TOG)*, volume 27, page 159. ACM, 2008. (Cited on page 6.)

[ST10] D Koc San and M Turker. Building extraction from high resolution satellite images using hough transform. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*, 38(Part 8), 2010. (Cited on page 18.)

[STD09] H. Sportouche, F. Tupin, and L. Denise. Building extraction and 3d reconstruction in urban areas from high-resolution optical and sar imagery. In *2009 Joint Urban Remote Sensing Event*, pages 1–11, May 2009. (Cited on page 28.)

[Sti75] G. Stiny. Pictorial and formal aspects of shape and shape grammars. In *Birkhauser Verlag*, 1975. (Cited on page 20.)

[SWRC09] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009. (Cited on page 15.)

[SZS03] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7):787–800, July 2003. (Cited on page 13.)

[TCGV09] Thoreau Rory Tooke, Nicholas C Coops, Nicholas R Goodwin, and James A Voogt. Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications. *Remote Sensing of Environment*, 113(2):398–407, 2009. (Cited on page 19.)

[The06] Lau Bee Theng. Automatic building extraction from satellite imagery. *Engineering Letters*, 2006. (Cited on page 18.)

[TJL88] M. Tuceryan, A. Jain, and Y. Lee. Texture segmentation using voronoi polygons. In *CVPR*, 1988. (Cited on page 43.)

[TK03] K. Tu-Ko. A hybrid road identification system using image processing techniques and backpropagation neural network. Mississippi State University, Starkville, 2003. (Cited on page 19.)

[TMT10] A. Toshev, P. Mordohai, and B. Taskar. Detecting and parsing architecture at city scale from range data. In *CVPR*, 2010. (Cited on page 22.)

[TMT16] A. Toshev, P. Mordohai, and B. Taskar. Neurally-guided procedural models: Learning to guide procedural models

with deep neural networks. In *29th Conference on Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016. (Cited on pages 22 and 96.)

[TQCR16] J. Tiana, R. Qinb, D. Cerraa, and P. Reinartza. Building change detection in very high resolution satellite stereo image time series. In *ISPRS*, 2016. (Cited on page 18.)

[TTM07] C. Tison, Florence Tupin, and H. Maître. A fusion scheme for joint retrieval of urban height map and classification from high resolution interferometric sar images. In *IEEE Transactions on Geoscience and remote Sensing*, volume 45, pages 495–505, February 2007. (Cited on page 16.)

[Tuy10] T. Tuytelaars. Dense interest points. In *CVPR*, 2010. (Cited on page 36.)

[TVG11] D. Tingdahl and L. Van Gool. A public system for image based 3d model generation. In *Computer Vision/Computer Graphics Collaboration Techniques*, page 262–273, 2011. (Cited on page 20.)

[TWZ08] Y. Taguchi, B. Wilburn, and C. Zitnick. Stereo reconstruction with mixed pixels using adaptive over-segmentation. In *CVPR*, 2008. (Cited on page 56.)

[TYK+12] J. Talton, L. Yang, R. Kumar, M. Lim, N. Goodman, and R. Mech. Learning design patterns with bayesian grammar induction. In *ACM SIGGRAPH*, 2012. (Cited on page 21.)

[VAB10] C. Vanegas, D. Aliaga, and B. Benes. Building reconstruction using manhattan-world grammars. In *CVPR*, 2010. (Cited on pages 20, 21 and 22.)

[VAM+09] C. Vanegas, D. Aliaga, P. Mueller, P. Waddell, B. Watson, and P. Wonka. Modeling the appearance and behavior of

urban spaces. In *EUROGRAPHICS State of the Art Reports*, 2009. (Cited on page 20.)

[VAM+10] C. Vanegas, D. Aliaga, P. Mueller, P. Waddell, B. Watson, and P. Wonka. Modeling the appearance and behavior of urban spaces. *Computer Graphics Forum*, 29(1):25–42, 2010. (Cited on page 20.)

[VCB10] S. Valero, J. Chanussut, and J. Benediktsson. Advanced directional mathematical morphology for the detection of the road network in very high resolution remote sensing images. *Pattern Recognition Letters*, 31(10):1120–1127, 2010. (Cited on page 19.)

[VdBBR+12] M. Van den Bergh, X. Boix, G. Roig, B. De Capitani, and L. Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *ECCV*, 2012. (Cited on pages 39 and 49.)

[Ver13] Yannick Verdie. *Urban scene modeling from airborne data*. Theses, Université Nice Sophia Antipolis, October 2013. Version éditée après soutenance (Novembre 2013). (Cited on page 10.)

[VGDA+12] A. Vanegas, A. Garcia-Dorado, D. Aliaga, B. Benes, and P. Waddell. Inverse design of urban procedural models. *ACM Transactions on Graphics (TOG)*, 31(6):168, 2012. (Cited on page 22.)

[VGJMR10] R. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *PAMI*, 32(4), 2010. (Cited on pages 39, 41, 51 and 91.)

[VKH06] V. Verma, R. Kumar, and S. Hsu. 3d building detection and modeling from aerial lidar data. In *CVPR*, 2006. (Cited on page 25.)

[VL14] Y. Verdie and F. Lafarge. Detecting parametric objects in large scenes by monte carlo sampling. *IJCV*, 106(1), 2014. (Cited on page 40.)

[VLA15] Y. Verdie, F. Lafarge, and P. Alliez. Lod generation for urban scenes. *ACM Transactions on Graphics*, 34(3), 2015. (Cited on pages 20 and 27.)

[VLPK12] H. Vu, P. Labatut, J. Pons, and R. Keriven. High accuracy and visibility-consistent dense multi-view stereo. *PAMI*, 34(5):889–901, 2012. (Cited on page 24.)

[VTC07] C. Vogiatzis, G.and Hernandez, P. Torr, and R. Cipolla. Multiview Stereo via Volumetric Graph-cuts and Occlusion Robust Photo-Consistency. *PAMI*, 29(12):2241–2246, 2007. (Cited on page 24.)

[WLH06] Oliver Wang, Suresh K. Lodha, and David P. Helmbold. A bayesian approach to building footprint extraction from aerial LIDAR data. In *3rd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT 2006), 14-16 June 2006, Chapel Hill, North Carolina, USA*, pages 192–199, 2006. (Cited on page 70.)

[WW12] J. Wang and X. Wang. Vcells: Simple and efficient superpixels using edge-weighted centroidal voronoi tessellations. *PAMI*, 34(6), 2012. (Cited on page 39.)

[WYD+14] F. Wu, D. Yan, W. Dong, X. Zhang, and P. Wonka. Inverse procedural modeling of facade layouts. In *ACM SIGGRAPH*, 2014. (Cited on page 21.)

[WYZ+16] Weixing Wang, Nan Yang, Yi Zhang, Fengping Wang, Ting Cao, and Patrik Eklund. A review of road extraction from remote sensing images. *Journal of Traffic and Transportation Engineering (English Edition)*, 3(3):271 – 282, 2016. (Cited on pages 19 and 31.)

[XDG14] G.-S. Xia, J. Delon, and Y. Gousseau. Accurate junction detection and characterization in natural images. volume 106, 2014. (Cited on pages 101 and 103.)

[XFZ+09] J. Xiao, T. Fang, P. Zhao, M. Lhuillier, and L. Quan. Image-based street-side city modeling. In *ACM SIGGRAPH Asia*, 2009. (Cited on page 21.)

[XGL16] X. Xiao, B. Guo, and D. Li. Multi-view stereo matching based on self-adaptive patch and image grouping for multiple unmanned aerial vehicle imagery. *Remote Sensing*, 8(2):89, 2016. (Cited on page 14.)

[XQ09] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 686–693, Sept 2009. (Cited on page 16.)

[YCZ+16] W. Yuan, S. Chen, Y. Zhang, J. Gong, and R. Shibasaki. An aerial-image dense matching approach based on optical flow field. In *ISPRS*, 2016. (Cited on page 14.)

[YHNF03] S. You, J. Hu, U. Neumann, and P. Fox. Urban site modeling from lidar. In *Part III. ICCSA*, 2003. (Cited on pages 25 and 26.)

[ZBKB08] L. Zebedin, J. Bauer, K.F. Karner, and H. Bischof. Fusion of feature- and area-based information for urban buildings

modeling from aerial imagery. In *ECCV*, 2008. (Cited on pages 18, 25 and 27.)

[ZFW+12] Z. Zhang, S. Fidler, J. Waggoner, Y. Cao, S. Dickinson, J. Siskind, and S. Wang. Superedge grouping for object localization by combining appearance and shape information. In *CVPR*, 2012. (Cited on pages 36 and 70.)

[ZGWX05] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu. What are textons? *IJCV*, 62(1/2), 2005. (Cited on page 39.)

[Zha04] C. Zhang. Towards an operational system for automated updating of road databases by integration of imagery and geodata. *Journal of Photogrammetry and Remote Sensing*, 58(3):166–168, 2004. (Cited on page 19.)

[ZK07a] C. Zitnick and S.B. Kang. Stereo for image-based rendering using image over-segmentation. *IJCV*, 75(1), 2007. (Cited on page 56.)

[ZK07b] L. Zitnick and S. B. Kang. Stereo for image-based rendering using image over-segmentation. *IJCV*, 75(1), 2007. (Cited on page 35.)

[ZN08] Q.Y. Zhou and U. Neumann. Fast and extensible building modeling from airborne lidar data. In *ACM GIS*, 2008. (Cited on pages 25 and 26.)

[ZN09] Q.Y. Zhou and U. Neumann. A streaming framework for seamless building reconstruction from large-scale aerial lidar data. In *CVPR*, 2009. (Cited on pages 25 and 26.)

[ZN10] Qian-Yi Zhou and Ulrich Neumann. 2.5d dual contouring: A robust approach to creating building models from aerial lidar point clouds. In *ECCV*. IEEE, 2010. (Cited on page 24.)

[ZN12] Qian-Yi Zhou and Ulrich Neumann. 2.5 d building modeling by discovering global regularities. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–333. IEEE, 2012. (Cited on page 25.)

[ZWDF15] E. Zheng, K. Wang, E. Dunn, and J.-M. Frahm. Minimal solvers for 3d geometry from satellite imagery. In *ICCV*, 2015. (Cited on pages 11 and 28.)

[ZWW+11] G. Zeng, P. Wang, J. Wang, R. Gan, and H. Zha. Structure-sensitive superpixels via geodesic distance. In *ICCV*, 2011. (Cited on page 39.)

[ZWY10] Chenxi Zhang, Liang Wang, and Ruigang Yang. Semantic segmentation of urban scenes using dense depth maps. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 708–721, Berlin, Heidelberg, 2010. Springer-Verlag. (Cited on page 16.)

[ZXJ+13] H. Zhang, K. Xu, W. Jiang, J. Lin, D. Cohen-Or, and B. Chen. Layered analysis of irregular facades via symmetry maximization. In *ACM SIGGRAPH*, 2013. (Cited on page 21.)